# Statistical tools for seed bank detection

**Jochen Blath**[1], **Eugenio Buzzoni**[2], **Jere Koskela**[3], **Maite Wilke-Berenguer**[4]

**Abstract**

We derive statistical tools to analyze the patterns of genetic variability produced by models related to seed banks; in particular the Kingman coalescent, its time-changed counterpart describing so-called weak seed banks, the strong seed bank coalescent, and the two-island structured coalescent. As (strong) seed banks stratify a population, we expect them to produce a signal comparable to population structure. We present tractable formulas for Wright's $F_{ST}$ and the expected site frequency spectrum for these models, and show that they can distinguish between some models for certain ranges of parameters. We then use pseudo-marginal MCMC to show that the full likelihood can reliably distinguish between all models in the presence of parameter uncertainty under moderate stratification, and point out statistical pitfalls arising from stratification that is either too strong or too weak. We further show that it is possible to infer parameters, and in particular determine whether mutation is taking place in the (strong) seed bank.

Keywords: seed bank, coalescent, population structure, model selection, site frequency spectrum, sampling formula.
2010 MSC: 92D10, 62P10.

[1]Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany; blath@math.tu-berlin.de

[2]Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany; buzzoni@tu-berlin.de

[3]Corresponding author. Department of Statistics, University of Warwick, Coventry CV4 7AL, UK; j.koskela@warwick.ac.uk

[4]Fakultät für Mathematik, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany; maite.wilkeberenguer@ruhr-uni-bochum.de

# 1 Introduction and basic models

## 1.1 Seed banks in population genetics

Seed banks, or reservoirs of dormant individuals that can be resuscitated in the future, are common in many communities of macroscopic (e.g. plant) and microscopic (e.g. bacterial) organisms. They extend the persistence of genotypes and are important for the diversity and functioning of populations. Microbial dormancy is common in a range of ecosystems, and there is evidence that the ecology and evolution of microbial communities are strongly influenced by seed banks. It has been observed that more that 90% of microbial biomass in soil is metabolically inactive. See [LJ11, SL18] for overviews on seed banks.

Seed banks have a significant influence on classical evolutionary forces such as selection and genetic drift. For example, seed banks can counteract the effect of genetic drift, and lead to population stratification. However, the development of a comprehensive population genetic theory incorporating seed banks is still in its early stages, and plenty of open questions remain [SL18]. While some basic mathematical models have been derived and predict unique patterns of genetic variability in idealized scenarios [LJ11, KKL01, ŽT12, BGE$^+$15, BGKW16, dHP17, KMTŽ17, HMTŽ18], statistical tools to infer the presence of 'weak' or 'strong' seed banks are still largely missing (however, see [SAMT19], which was produced in parallel with this work).

The aim of this article is to provide basic statistical tools to analyze patterns of genetic variability produced by the above models of seed banks. We also assess the utility of these tools for parameter estimation and model selection based on genetic data. Notably, we will provide comparisons between variability under seed banks, and classical models of population structure [Her94]. Both model classes can be expected to predict somewhat similar patterns of diversity, and we will study the extent to which sequence data can differentiate between. This is motivated by the need to understand the roles of dormancy and biogeography in microbial communities [LJ11, p.125]. We extend earlier studies [TLL$^+$11, BGE$^+$15], where seed banks were compared to panmictic models. We begin with a brief review of the relevant genetic models with and without seed banks.

## 1.2 Population models

*Kingman's coalescent* (K): The standard model of genetic ancestry in the absence of a seed bank is the *coalescent* (or *Kingman's coalescent*) [Kin82], which describes ancestries of samples of size $n \in \mathbb{N}$ from a large, selectively neutral, panmictic population of size $N \gg n$ following e.g. a Wright-Fisher model. Measuring time in units of $N$ and tracing the ancestry of a sample of size $n \ll N$ backwards in time results in a coalescent process $\Pi^n$ in which each pair of lineages merges to a common ancestor independently at rate 1 as $N \to \infty$. A rooted ancestral tree is formed once the most recent common ancestor of the whole sample is reached. We denote this scenario by `K`. This model is currently the standard null model in population genetics (see e.g. [Wak09] for an introduction) and arises from a large class of population models.

*'Weak' seed banks and the delayed coalescent* (`W`): The coalescent was extended in [KKL01] to incorporate a *'weak' seed bank*. In this model, an individual inherits its genetic material from a parent that was alive a random number of generations ago. The random separation is assumed to have mean $\beta^{-1}$ for some $\beta \in (0, 1]$. Measuring time in units of $N$ and tracing the ancestry of a sample of size $n \ll N$ as above, it can be shown that the genealogy is still given by a coalescent in which each pair of lineages merges to a common ancestor independently with rate $\beta^2$. Thus, the effect of the seed bank is to stretch the branches of the Kingman coalescent by a constant factor [KKL01, BGKS13], but the topology and relative branch lengths remain identical to those of the coalescent. Thus the weak seed bank coalescent with mean separation $\beta^{-1}$ and population-rescaled mutation rate $u > 0$ is statistically identical to Kingman's coalescent with population-rescaled mutation rate $u/\beta^2$, and e.g. the normalized site frequency spectrum under the infinitely many sites model is invariant between these models [BGE$^+$15]. Nevertheless, the seed bank does have important consequences e.g. for the estimation of effective population size and mutation rates in the presence of prior information, or some other means of resolving the lack of identifiability. We call the corresponding coalescent a 'delayed coalescent' and denote this scenario by `W`.

*'Strong' seed banks and the seed bank coalescent* (`S`): The recent model in [BGKW16] extends the Wright Fisher framework to a model with a classical 'active' population of size $N$ and a separate 'seed bank' of comparable size $M := \lfloor N/K \rfloor$, for some $K > 0$, allowing for 'migration' of a fraction of $\lfloor c/N \rfloor$ individuals between the two subpopulations. The active population follows a Wright-Fisher model, while the dormant population in the seed bank persists without reproducing. This model can be seen as a mathematical formalization of [LJ11, Figure 2]. The age structure in the resulting seed bank is geometric with mean of order $N$, which means that seeds can remain viable in the seed bank for $O(N)$ generations. Measuring time in units of $N$, the genealogy of a sample of size $n^{(1)} \ll N$ (resp. $n^{(2)} \ll M$) from the active (resp. dormant) population, is described by the so-called *seed bank coalescent* [BGKW16], in which active lineages fall dormant at rate $c$ and coalesce at rate 1 per pair, while dormant lines resuscitate at rate $cK$. We call this ancestral process a *(strong) seed bank coalescent*, and denote this scenario by `S`. The seed bank coalescent has a very different site frequency spectrum to the classical and weak seed bank coalescents [BGE$^+$15].

*The two island model and the structured coalescent (*`TI`*)*: Having modeled a strong seed bank as a separate population linked to the active one via migration, it is natural investigate its relation to Wright's two island model [Her94, Wak09]. In the simplest case (which we assume throughout) there are two populations (1 and 2) of respective sizes $N$ and $M = \lfloor N/K \rfloor$, with a fixed fraction of $\lfloor c/N \rfloor$ individuals migrating both from 1 to 2 and from 2 to 1 each generation. Measuring time in units of $N \to \infty$ generations, the genealogy of a sample of respective sizes $n^{(1)} \ll N$ and $n^{(2)} \ll M$ from islands 1 and 2 is described by a similar ancestral process as the strong seed bank coalescent, except that pairs of lineages in population 2 also merge independently with rate $1/K$. We denote this scenario by `TI`. The resulting ancestral process is the *structured coalescent* [Her94, Not90], which describes the ancestry of a geographically structured population with migration.

In this article we investigate the extent to which genetic data can distinguish between models `K`, `W`, `S`, and `TI`. All four are a priori plausible as models for various real populations. In [TLL$^+$11], the authors studied two species of wild tomato (S. chilense and S. peruvianum), and inferred average seed bank delays of 9 and 12 generations. Estimates of corresponding effective population sizes are $O(10^5)$ [ASS07], which suggests that scenario `W` is appropriate. On the other hand, dormant bacteria have been observed to remain viable for millions of years [VRP00], which suggests that the strong seed bank could be relevant. A stable reservoir of dormant individuals requires periods of dormancy on the order of the effective population size [BGE$^+$15], so that model `S` seems appropriate whenever there is a stable reservoir of dormant types, with individuals switching between reservoirs with some fixed rate as outlined in [LJ11] for bacterial communities. These considerations highlight the need to distinguish the two types of seed banks from data in cases where the presence or size of a seed bank or the typical period of dormancy are uncertain. It is also of interest to distinguish the signal of (strong) seed banks from geographic structure, which could in principle produce similar patterns of genetic stratification in the population.

### 1.3 Mutation models and key statistical quantities

We consider three models of genetic diversity and mutation: the infinite alleles model (IAM), the infinite sites model (ISM), and the finite alleles model (FAM). The FAM is in less widespread use due to its high computational demands, so we postpone results under it to the appendix. We also only present results under the FAM for the special case of two alleles, but our work generalizes to any number.

We consider several classical statistical quantities: the sample heterozygosity and Wright's $F_{ST}$ [Wri51], the site frequency spectrum (SFS), and the sampling distribution of the full sequence data. These measures are informative about the underlying coalescent scenario, and suited to the different mutation models, to varying degrees. They also differ in the extent to which they are tractable. The sample heterozygosity, Wright's $F_{ST}$ and the (normalized) SFS discard statistical signal, but are readily computed (at least numerically) in most settings. The sampling distribution of the sequence data fully captures the signal in a data set, but is available only via Monte Carlo schemes. Our results clarify when computationally cheap summary statistics suffice to distinguish between models, and when the full likelihood is needed.

*The infinite alleles model (IAM)*: Given a coalescent tree distributed according to any of the models introduced above, a sample of genetic data from the infinite alleles model is generated by assigning an arbitrary allele to the most recent common ancestor, and simulating mutations along the branches of the coalescent tree with population-rescaled mutation rate $u := N\mu > 0$ for the branches in the first (and possibly only) population and $u' := M\mu' \geqslant 0$ in the second population (if one is present). Above, $N$ and $M$ represent effective population sizes, while $\mu$ and $\mu'$ are the per-site, per-generation mutation probabilities. Each mutation results in a new parent-independent allele that has never existed in the population before, and alleles are inherited along lineages. Population-rescaled mutation rates in further mutation models below are defined analogously.

We encode a sample of size $n^{(1)} + n^{(2)} = n$, where $n^{(i)}$ is the sample size from population $i$, as the pair of $n$-tuples $(\mathbf{n}^{(1)}, \mathbf{n}^{(2)})$, where $n_j^{(i)}$ is the number of $j$ alleles on island $i$ under some fixed but arbitrary ordering of observed alleles, and $n^{(i)} = \sum_j n_j^{(i)}$. Both tuples are padded by zeros if fewer than $n$ distinct alleles are observed for notational convenience, and we will drop the superscripts and second tuple for models with only one population.

The (somewhat out-dated) infinite alleles model is appropriate when the data only encodes when two alleles are different, but no further detail is available, such as is the case for electrophoresis data [HL66].

*The infinite sites model (ISM)*: We now identify the locus with the unit interval $[0, 1]$. Mutations, which continue to occur on the branches of the coalescent tree with rates $u$ and $u'$, are assumed to occur at distinct sites on the locus, and are inherited along the branches of the tree so that the allele of an individual is the list of all mutations along its ancestral line. Thus, the whole history of mutations up to the root is retained. A sample of size $n := n^{(1)} + n^{(2)}$ is specified by the triple $(\mathbf{t}, \mathbf{n}^{(1)}, \mathbf{n}^{(2)})$, where $\mathbf{t} := (t_1, \ldots, t_d)$ is the list of all observed alleles, and $n_j^{(i)}$ is the observed number of allele $t_j$ in population $i$. For details on this parametrization of the infinite sites model and its relation to coalescent models see e.g. [BB08].

*The finite alleles model (FAM)*: We consider a finite set of possible allele identified with $\{1, \ldots, d\}$. The type of the most recent common ancestor is sampled from some probability mass function $\rho = (\rho_1, \ldots, \rho_d)$, and mutations occur along the branches of the coalescent tree at rates $u$ and $u'$ as before. At a mutation, a new allele is sampled from a $d \times d$ stochastic matrix $P$, and alleles are inherited along branches as before. A sample under the FAM is also described by the pair of tuples $(\mathbf{n}^{(1)}, \mathbf{n}^{(2)})$, with the distinction that each tuple is now of fixed length $d$. Throughout the article, we take $d = 2$, and set $u_2 := uP_{12}$ as well as $u_1 := uP_{21}$ for notational brevity, and define mutation rates $u_1'$ and $u_2'$ for a second population analogously.

Note that the classical Watterson estimator of mutation rate depends on the chosen coalescent model. Further, in scenarios `TI` and `S`, we will allow the overall mutation rate to differ between active and dormant lineages. Determining whether mutations take place on dormant lineages in nature, perhaps at a reduced rate, is an interesting open question [SL18], and one of our motivations was to determine whether it is answerable from DNA sequence data.

### 1.4 Diffusion models

All four coalescent models are dual to their respective *Wright-Fisher diffusions*, the exact form of which depends on the accompanying mutation model. The FAM, `TI` Wright-Fisher diffusion solves the pair of SDEs

$$\begin{aligned} \mathrm{d}X(t) &= [u_2(1 - X(t)) - u_1 X(t) + c(Y(t) - X(t))]\mathrm{d}t \\ &\quad + \alpha\sqrt{X(t)(1 - X(t))}\mathrm{d}B(t), \\ \mathrm{d}Y(t) &= [u_2'(1 - Y(t)) - u_1' Y(t) + Kc(X(t) - Y(t))]\mathrm{d}t \\ &\quad + \alpha'\sqrt{Y(t)(1 - Y(t))}\mathrm{d}B'(t), \end{aligned} \tag{1}$$

with initial value $(X(0), Y(0)) = (x, y) \in [0,1]^2$, where $1/\alpha^2$, $1/(\alpha')^2$ are effective population sizes, and $\{B_t\}$, $\{B'_t\}$ are independent Brownian motions. Duals to scenarios `K`, `W`, and `S` can be recovered as special cases: for `K` we set $\alpha = 1$ and $c = 0$, for `W` we take $\alpha = \beta$ and $c = 0$, and for `S` we take $\alpha = 1$ and $\alpha' = 0$. For scenarios `K` and `W` we also only consider the $X$-coordinate, and in scenario `S`, the $X$-coordinate corresponds to the active population, while $Y$ is the seed bank. In each case the solution is an ergodic diffusion with a unique stationary distribution on $[0,1]$ (or $[0,1]^2$), which we will denote by $\mu^{\mathtt{I}}$ for $I \in \{\mathtt{K}, \mathtt{W}, \mathtt{S}, \mathtt{TI}\}$. It is also possible to derive the analogue of the Wright-Fisher diffusion for the IAM and ISM. This leads to measure-valued diffusions, or *Fleming-Viot processes* [EK86], which we do not require in our analysis.

### 1.5 Outline of the paper

In Section 2 we discuss Wright's $F_{ST}$ and the site frequency spectrum (SFS). We use phase-type distribution methods [HSJB19] to compute the expected SFS, and show that these statistics can distinguish between our scenarios to some extent. Since they are cheap to compute, they serve as a plausibility check for the presence of seed banks. Results for Wright's $F_{ST}$ for the FAM are presented in the appendix.

In Section 3 we present recursions for the likelihood functions of observed sequence data for the IAM and ISM under scenario `S`, which are currently missing in the literature. The recursions are intractable for large sample sizes, so we provide low-variance importance sampling schemes to approximate their solutions. Corresponding results for the FAM are presented in the appendix.

In Section 4 we provide statistical machinery for model selection and parameter inference for all scenarios under the ISM, which is the most relevant for handling of real data. We employ a pseudo-marginal Metropolis-Hastings algorithm for simultaneous model selection and parameter inference for the different models and assess its effectiveness with simulated data sets. We also address the specific question of detecting mutation in the (strong) seed bank.

We conclude the paper with a discussion of our results in Section 5.

## 2 Classical measures of population structure

In this section we investigate classical summary statistics for inferring population structure, namely Wright's $F_{ST}$ and the (normalized) site frequency spectrum nSFS. Unless stated otherwise, we assume positive mutation rates in all (sub-)populations.

### 2.1 Wright's $F_{ST}$ for seed banks and structured populations

Wright's $F_{ST}$ [Wri51] is a prominent but crude measure for population structure. There are various (more-or-less equivalent) formulations in the literature. Here, we follow the notation and interpretation of Herbots [Her94, p. 73] (see also [Rou04,

Chapter 3]), which studies this quantity for various structured models. Define

$$F_{ST} := \frac{p_0 - \bar{p}}{1 - \bar{p}}, \tag{2}$$

where $\bar{p}$ is the probability of *identity* of two genes sampled uniformly at random from the whole population, while $p_0$ is the probability of identity of two genes sampled uniformly from a single sub-population, itself previously randomly sampled with probability given by its relative population size.

For the FAM, $\bar{p}$ and $p_0$ are determined by the *sample homozygosity* (discussed in the appendix), whereas for the IAM and ISM, they are given in terms of *identity by descent*. Positive values of $F_{ST}$ indicate population structure, though its exact interpretation depends on the biological scenario. Hartl and Clark argue that $F_{ST} \in (0.05, 0.15)$ constitutes "moderate" genetic differentiation [HC97, Section 6.2], though applying such a rule indiscriminately can be misleading as $F_{ST}$ values depend on e.g. the life cycle and reproductive characteristics of the species, as well as on the details of spatial structure. We will be interested how the quantity compares between `S` and `TI`, where the latter certainly represents a strongly structured population.

**Wright's $F_{ST}$ for the IAM** Under the IAM, every mutation leads to a distinct allele. Hence, two sampled individuals are identical if and only if neither of their ancestral lineages mutated since the time of their most recent ancestor. Thus $p_0$ and $\bar{p}$ from (2) can be expressed as the so-called probabilities of *identity by descent* (IBD), and these probabilities can easily be represented in terms of the relevant coalescent.

Let $T$ be the (random) *time to the most recent common ancestor* (TMRCA) of a sample of size 2 in any of the above coalescent models and observe that, if we assume the same mutation rate $u = u'$ in both sub-populations (for $\mathtt{S}, \mathtt{TI}$), the probability that we do not see any mutations along the branches of the coalescent up to a time $t > 0$ is given by $e^{-2ut}$. Since mutations occur conditionally independently given $T$, we have

$$p_0 = \mathbb{E}_{\pi_0}[e^{-2uT}] \qquad \text{and} \qquad \bar{p} = \mathbb{E}_{\bar{\pi}}[e^{-2uT}],$$

where

$$\pi_0 := \left( \frac{K}{1+K}, 0, \frac{1}{1+K}, \right), \qquad \bar{\pi} := \left( \frac{K^2}{(1+K)^2}, \frac{2K}{(1+K)^2}, \frac{1}{(1+K)^2} \right).$$

In words, $\mathbb{E}_{\pi_0}$ is the expectation when the both genes are sampled from the same population, itself previously sampled among all populations according to its relative size, and $\mathbb{E}_{\bar{\pi}}$ is the expectation when the genes are sampled uniformly from the whole population. IBD has recently been investigated for `S` in [dHP17] in the case of a finite population with seed bank on a discrete torus.

To obtain an expression for IBD for distinct mutation rates $u \neq u'$, we need to trace the time the lineages spend in each population before the TMRCA. Let $R_{2,0}$, $R_{1,1}$ and $R_{0,2}$ be the time until coalescence the ancestral lineages spend both in the

first population, one lineage in each population and both in the second population, respectively. Then $T = R_{2,0} + R_{1,1} + R_{0,2}$ and we get

$$p_0 = \mathbb{E}_{\pi_0}\left[e^{-2uR_{2,0}-(u+u')R_{1,1}-2u'R_{0,2}}\right],$$
$$\bar{p} = \mathbb{E}_{\bar{\pi}}\left[e^{-2uR_{2,0}-(u+u')R_{1,1}-2u'R_{0,2}}\right].$$

Phase-type distribution theory [HSJB19] yields elegant closed form expressions for these quantities.

**Proposition 2.1.** *Assuming the IAM, the fixation index $F_{ST}^{\mathtt{I}}$ for $\mathtt{I} \in \{\mathtt{S}, \mathtt{TI}\}$ is given by*

$$F_{ST}^{\mathtt{I}} = \frac{p_0^{\mathtt{I}} - \bar{p}^{\mathtt{I}}}{1 - \bar{p}^{\mathtt{I}}}$$

*where*

$$p_0^{\mathtt{I}} = \pi_0(A - S^{\mathtt{I}})^{-1}s^{\mathtt{I}} \qquad \text{and} \qquad \bar{p}^{\mathtt{I}} = \bar{\pi}(A - S^{\mathtt{I}})^{-1}s^{\mathtt{I}}$$

*where $A$ is a diagonal matrix with diagonal $[-2u, -(u+u'), -2u']$, and*

$$S^{\mathtt{I}} = \begin{bmatrix} -(2c+1) & 2c & 0 \\ cK & -(cK+c) & c \\ 0 & 2cK & -(2cK+\alpha^{\mathtt{I}}) \end{bmatrix} \quad \text{and} \quad s^{\mathtt{I}} = \begin{bmatrix} 1 \\ 0 \\ \alpha^{\mathtt{I}} \end{bmatrix},$$

*where $\alpha^{\mathtt{S}} = 0$ and $\alpha^{\mathtt{TI}} = 1/K$.*

The proof is obtained using the machinery of [HSJB19] and we adhere to the notation used therein for the convenience of the reader. See [HSJB19, Example 2.4] for some different functionals of the seed bank coalescent obtained in this way.

*Proof.* Let $\{Z_t\}$ be a time-continuous Markov chain on the finite space

$$E_2 := \{(2,0), (1,1), (0,2), (*,*)\}$$

with Q-matrix

$$Q^{\mathtt{I}} = \begin{bmatrix} S^{\mathtt{I}} & s^{\mathtt{I}} \\ 0 & 0 \end{bmatrix}$$

for $\mathtt{I} \in \{\mathtt{S}, \mathtt{TI}\}$. For each model, $\{Z_t\}$ traces whether the lineages of a sample of 2 are both in the first population, one in each population or both in the second population. The state $(*,*)$ is reached at time $T$, and is absorbing.

Recall that $R_{2,0}$ was the time the ancestral lineages of the sample spent both in the first population and note that we can write it as

$$R_{2,0} = \int_0^T \mathbb{1}_{\{(2,0)\}}(Z_t)\mathrm{d}t.$$

We can do the same for $R_{1,1}$ and $R_{0,2}$, and thus [HSJB19, Theorem 2.5] yields

$$\begin{aligned} p_0 &= \mathbb{E}_{\pi_0}\left[e^{-2uR_{2,0}-(u+u')R_{1,1}-2u'R_{0,2}}\right] \\ &= \pi_0\left(\begin{bmatrix} -2u & 0 & 0 \\ 0 & -(u+u') & 0 \\ 0 & 0 & -2u' \end{bmatrix} - S^{\mathtt{I}}\right)^{-1} s^{\mathtt{I}} \end{aligned}$$

and analogously for $\bar{p}$. □

Figure 1 illustrates the $F_{ST}$ under different choices of parameters for the IAM. The pictures differ only slightly from those of the FAM in Figure 5 in the appendix.

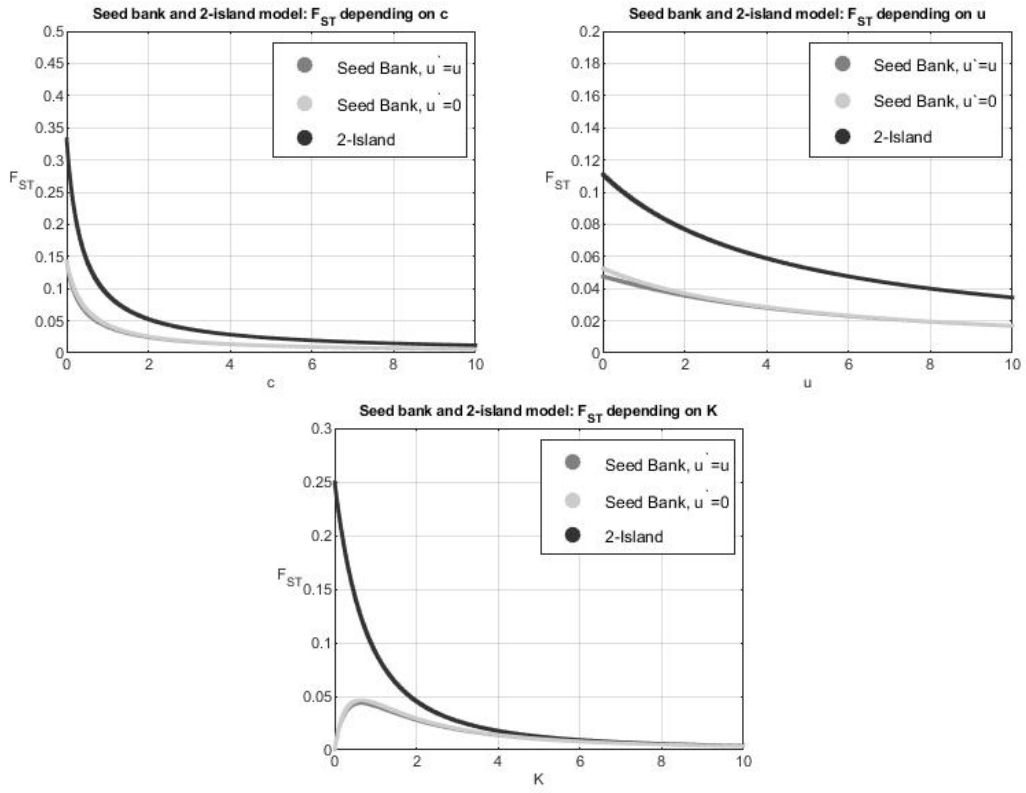



Figure 1: $F_{ST}$ under `S` and `TI` as a function of various parameters in the IAM. Where not specified, $K = c = 1$, $u_1 = u_2 = 0.5$.

**Wright's $F_{ST}$ for the ISM** The central difference between the IAM and the ISM is that all previous mutations on a lineage remain observable in the latter. However, this does not affect the probability of IBD of two individuals — they will still carry the same allele if and only if neither ancestral line mutated between the TMRCA and the present. Thus, sample heterozygosity $H$ and $F_{ST}$ under the ISM can be computed in exactly the same way as in the IAM.

## 2.2 The site frequency spectrum (SFS) in the ISM

The SFS is one of the most frequently used summary statistics under the ISM. For a sample of size $k$ it is given by a vector $(\zeta_1^{(k)}, \dots, \zeta_{k-1}^{(k)})$, with $\zeta_i^{(k)}$ denoting the number of sites at which the *derived* allele is observed $i$ times in the sample. This assumes that we know the wildtype and are therefore able to determine which of the two alleles is derived, and which is ancestral. In the case where we do not know which allele is which, the *folded* SFS $(\eta_1^{(k)}, \dots, \eta_{\lfloor k/2 \rfloor}^{(k)})$ can be used instead, where $\eta_i^{(k)}$ is the number of sites where two alleles are observed with multiplicities $i : k - i$.

The SFS is well understood for the classical Kingman coalescent `K`, and thus also in the case `W`, since the weak seed bank coalescent is just a constant time-change of the Kingman coalescent [ŽT12, Formula 1].

We can also calculate the *expected* SFS for the cases `TI` and `S`. We consider $k$ individuals sampled according to some initial distribution $\pi$ from the first and the second population. Since mutations in the ISM occur according to a Poisson process conditionally on the coalescent, $\mathbb{E}_\pi[\zeta_i^{(k)}]$ is the product of the mutation rate and the expected total lengths of branches that are ancestral to $i$ individuals, for which phase-type distribution theory is well suited. In order to state the result (and thereby give the bulk of the proof), we require a few technical definitions, but the calculation of the SFS then reduces to a simple vector-matrix multiplication in Proposition 2.2. The structure is reminiscent of the observations for the SFS of $\Lambda$-coalescents in [HSJB19].

As in Proposition 2.1 we want to define an auxiliary Markov chain. Its state space $E$ should be small to minimize computational cost, but needs to be sufficiently large to contain all information necessary to calculate the SFS, i.e. we need to know how many lineages are ancestral to $i$ individuals in the sample at any time in the coalescent, and how many of these lineages are in the first and second populations, respectively, in order to account for different mutation rates. For a sample of size $k$ define

$$E := \left\{ a \in \{0, \dots, k\}^{2k} \;\middle|\; \sum_{i=1}^{k} i(a_i + a_{k+i}) = k \right\} \setminus \{e_k, e_{2k}\}$$

where $e_k$ and $e_{2k}$ are the vectors with the entry 1 in positions $k$ and $2k$ respectively (and thus 0 everywhere else). We remove these in order to identify them as what will be the unique absorbing state of the Markov chain. Thus define

$$E^* := E \cup \{*\}.$$

For $a \in E$, if $i = 1, \dots k$, the quantity $a_i$ is the number of lineages currently in the first population that are ancestral to $i$ of the sampled individuals (independently of their origin). If $i = k+1, \dots, 2k$ then $a_i$ is the analogous number of lineages currently in the second population.

Given this interpretation, it becomes easy to identify the set $E_0$ of sensible starting points for the auxiliary Markov chain:

$$E_0 := \{a \in E \mid a_1 + a_{k+1} = k\}.$$

Starting in $a \in E_0$ corresponds to a sample of $a_1$ individuals from the first and $a_{k+1}$ individuals from the second population. Let $\pi$ be the initial a distribution of the Markov chain, assumed concentrated on $E_0$.

The only allowed transitions of the chain will be those corresponding to a coalescence or a migration. For $z \in \mathbb{Z}$ let $(z)^+ := \max\{z, 0\}$ and $(z)^- := \min\{z, 0\}$. We call a transition from the state $a \in E$ to $b \in E$ a *coalescence* if

(i) $\sum_{j=1}^{2k}(b_j - a_j)^- = -2$,

(ii) $\sum_{j=1}^{2k}(b_j - a_j)^+ = 1$,

(iii) $\sum_{j=1}^{k} j(b_j - a_j) = 0$.

The first two describe the effect of the coalescence of two lineages. The last sum only runs until $k$, ensuring that the coalescence takes place between lineages in the same population. A transition from $a$ to $b$ will be called a *migration* if

(i) $\sum_{j=1}^{2k}(b_j - a_j)^- = -1$,

(ii) $\sum_{j=1}^{2k}(b_j - a_j)^+ = 1$.

The rates at which the Markov chain then transitions between the states $a, b \in E$ depend on the model and are given by

$$S^{\mathtt{I},c}_{a,b} := \prod_{j=1:\, b_j - a_j < 0}^{k} \binom{a_j}{b_j - a_j} + \alpha^{\mathtt{I}} \prod_{j=1:\, b_{k+j} - a_{k+j} < 0}^{k} \binom{a_{k+j}}{b_{k+j} - a_{k+j}},$$

if $a \mapsto b$ is a *coalescence* and

$$S^{\mathtt{I},m}_{a,b} := c \sum_{j=1:\, b_j - a_j < 0}^{k} a_j + cK \sum_{j=1:\, b_{k+j} - a_{k+j} < 0}^{k} a_{k+j},$$

if it is a *migration*, where we again set $\alpha^{\mathtt{I}} = 0$ if $\mathtt{I} = \mathtt{S}$ and $\alpha^{\mathtt{I}} = 1/K$ if $\mathtt{I} = \mathtt{TI}$.

Next, define $s^{\mathtt{I}} : E \to [0, \infty[$ as

$$s^{\mathtt{I}}(a) := \begin{cases} 1, & \text{if } \sum_{j=1}^{2k} a_j = \sum_{j=1}^{k} a_j = 2, \\ \alpha^{\mathtt{I}}, & \text{if } \sum_{j=1}^{2k} a_j = \sum_{j=k+1}^{2k} a_j = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $s^{\mathtt{I}}$ is non-zero precisely on the states with two lineages remaining which could coalesce into the absorbing state $*$, and gives the rate of that event.

With this now define the matrix $S^{\mathtt{I}} = (S^{\mathtt{I}}_{a,b})_{a,b \in E}$ through

$$S^{\mathtt{I}}_{a,b} := \begin{cases} S^{\mathtt{I},c}_{a,b}, & \text{if } a \mapsto b \text{ is a coalescence,} \\ S^{\mathtt{I},m}_{a,b}, & \text{if } a \mapsto b \text{ is a migration,} \\ -s^{\mathtt{I}}(a) - \sum_{a' \neq a} S_{a,a'} & \text{if } a = b, \\ 0, & \text{otherwise .} \end{cases}$$

Finally, we define $r_i(*) := 0$ for any $i = 1, \dots, k-1$, and for every $a \in E$,

$$r_i(a) := ua_i + u'a_{k+i}.$$

The vectors $\pi$, $r_1, \dots, r_{k-1}$ can be taken as normal vectors by fixing an ordering on $E^*$, which also justifies representing $S^{\mathtt{I}}$ as a matrix. Hence (3) should be read as a vector-matrix multiplication.

**Proposition 2.2.** *Assume the ISM, with mutation rates $u, u' \geqslant 0$ in the first and second population, respectively. Let $\pi$ describe how the $k \in \mathbb{N}$ individuals are sampled from the first and second population. Then*

$$\mathbb{E}_\pi\left[\zeta_i^{(k)}\right] = \pi(-S^{\mathtt{I}})^{-1} r_i \tag{3}$$

*for all $i = 1, \dots, k-1$ and $\mathtt{I} \in \{\mathtt{TI}, \mathtt{S}\}$.*

For a sample of $k_1$ individuals from the first population and $k_2 = k - k_1$ individuals from the second population, set $\pi = \pi^{(k_1,k_2)} := \delta_{(k_1,0\dots,0,k_2,0,\dots,0)}$, where the right hand side is the Dirac delta measure and the non-zero entries are in positions 1 and $k+1$. For a sample drawn uniformly from the whole population, set $\pi(a) = \pi^{\mathrm{unif}}(a) := \binom{k}{a_{k+1}} K^{a_{k+1}} (K+1)^{-k}$ for any $a \in E_0$.

*Proof.* Let $\{Z_t\}$ be a Markov process with state space $E^*$ and Q-matrix

$$Q := \begin{bmatrix} S^{\mathtt{I}} & s^{\mathtt{I}} \\ 0 & 0 \end{bmatrix}.$$

Started in $\pi$, the time $\{Z_t\}$ absorbs into $*$ is equal in distribution to the time to the most recent common ancestor of a sample of size $k$ drawn according to $\pi$. Since mutations occur independently of the coalescent given the ancestry, to compute $\mathbb{E}_\pi[\zeta_i^{(k)}]$ we trace the time a lineage in the coalescent is ancestral to $i$ of the initial individuals and multiply it by $u$ when it is in the first and by $u'$ when it is in the second population. This is done by defining

$$\tilde{\zeta}_i^{(k)} := \int_0^\tau r_i(Z_t) \mathrm{d}t,$$

and noting that

$$\mathbb{E}_\pi\left[\zeta_i^{(k)}\right] = \mathbb{E}_\pi\left[\tilde{\zeta}_i^{(k)}\right].$$

Thus, [HSJB19, Eq (8)] yields (3) above. □

**Remark 2.3.** The normalized expected site frequency spectrum [EBBF15, p. 13] (NESFS) $(E\hat{\zeta}_1^{(k)}, \dots, E\hat{\zeta}_{k-1}^{(k)})$ is defined as

$$E\hat{\zeta}_i^{(k)} := \frac{\mathbb{E}[\zeta_i^{(k)}]}{\sum_{l=2}^k l\mathbb{E}[T_l]},$$

where $T_l$ is the time during which there are $l$ distinct lineages in the coalescent regardless of to which population they belong. In other words, $\sum_{l=2}^{k} l\mathbb{E}[T_l]$ is the average tree length. The NESFS is a first-order approximation of the expectation of the normalized SFS [EBBF15, p. 9], given by

$$\hat{\zeta}_i^{(k)} := \frac{\zeta_i^{(k)}}{\zeta_1^{(k)} + \cdots + \zeta_{k-1}^{(k)}}.$$

The distribution of $(\hat{\zeta}_1^{(k)}, \ldots, \hat{\zeta}_{k-1}^{(k)})$ is very insensitive to the mutation rate, provided it is not too small, facilitating practical inference when the mutation rate is unknown [EBBF15, Supporting Information, pages SI12 – SI13]. The average tree length for `S` was analyzed in [HSJB19] and thus all necessary quantities to calculate the normalized expected SFS similarly to the SFS are given.

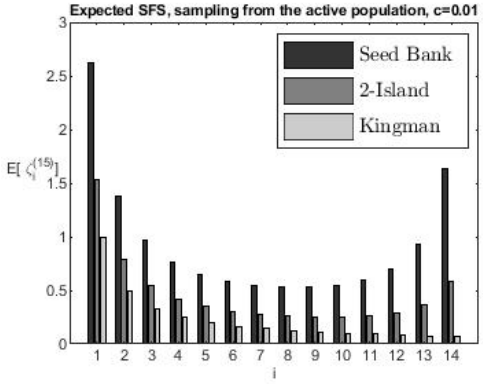
Figures 2 and 3 provide illustrations of the expected SFS, with and without normalization. It is noteworthy that the magnitude of entries in the expected SFS varies strongly between the three models, while `S` and `TI` have very similar normalized spectra when only sampling one (in the case of `S`, active) population. The implication is that all three models are straightforward to tell apart if the population-rescaled mutation rate is known, but that a larger sample or a more informative statistic is needed to distinguish `S` from `TI` when it is unknown. When dormant lineages are included in the sample, `S` predicts an excess of singletons, which can be detected even from the nSFS. The excess of singletons is due to the fact that only few distinct active lineages will remain in the genealogy by the time that the ancestral line of the dormant samples first fell dormant. Thus, the external branch connecting a dormant lineage to the rest of the ancetral tree is likely to be much longer than the external branches between active samples.

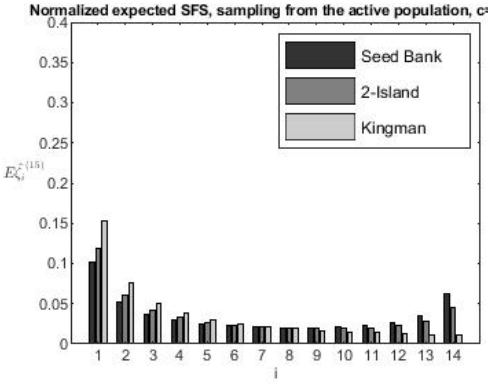
# 3 Recursions for the sampling distributions

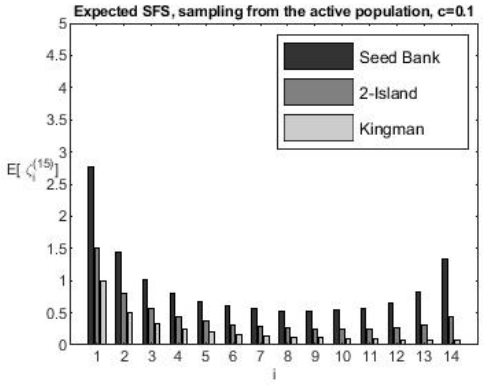
In this section we use recursions to characterize the (in general intractable) sampling distributions for scenario `S` under the IAM and ISM. Similar results under the FAM are provided in the appendix. The corresponding recursions for `K`, `W`, and `TI` are special cases of [DG04, Eq (2)]. We will also describe a low-variance Monte Carlo scheme to approximate solutions of these recursions, and hence conduct unbiased inference and model selection based on full likelihoods.

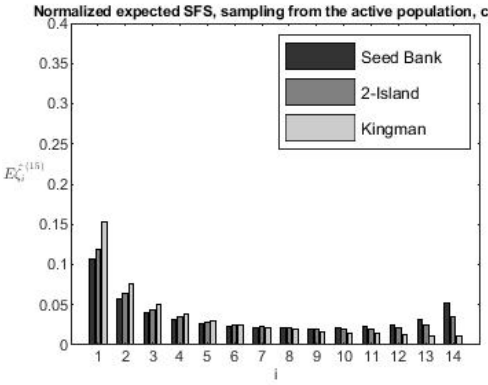
## 3.1 IAM recursion

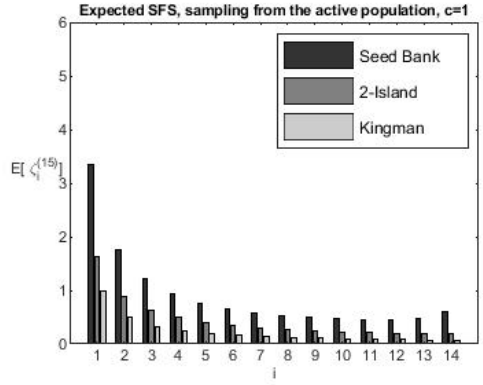
Let $p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)})$ be the probability of observing sample $\mathbf{n}^{(1)}$ from the active population, and $\mathbf{n}^{(2)}$ from the seed bank under `S`, and $\mathbf{e}_i$ be the canonical unit vector with

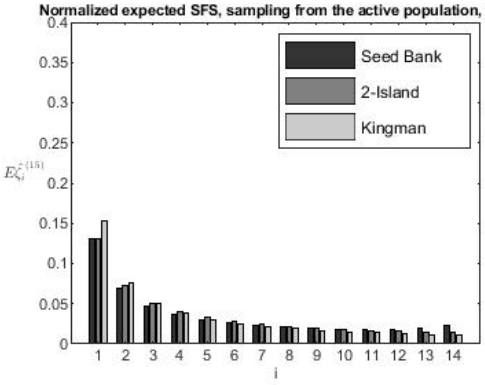

Expected SFS, sampling from the active population, c=0.01
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i

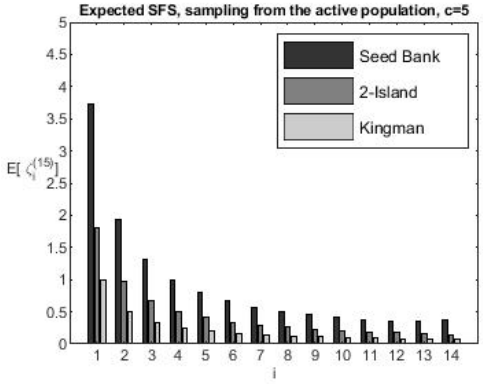

Normalized expected SFS, sampling from the active population, c=0.01
Seed Bank
2-Island
Kingman
i

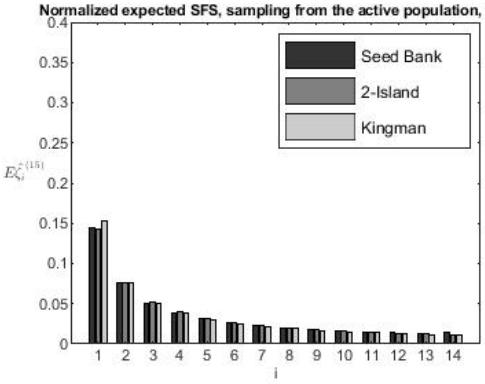

Expected SFS, sampling from the active population, c=0.1
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i


Normalized expected SFS, sampling from the active population, c=0.1
Seed Bank
2-Island
Kingman
i


Expected SFS, sampling from the active population, c=1
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i


Normalized expected SFS, sampling from the active population, c=1
Seed Bank
2-Island
Kingman
i


Expected SFS, sampling from the active population, c=5
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i


Normalized expected SFS, sampling from the active population, c=5
Seed Bank
2-Island
Kingman
i

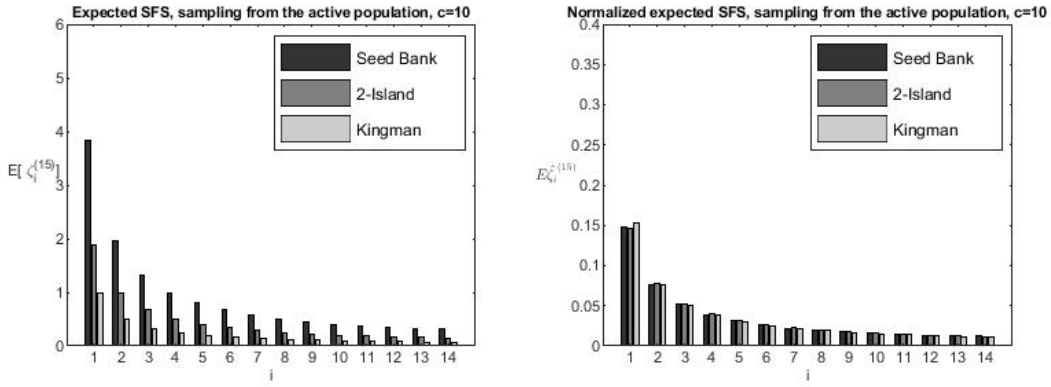
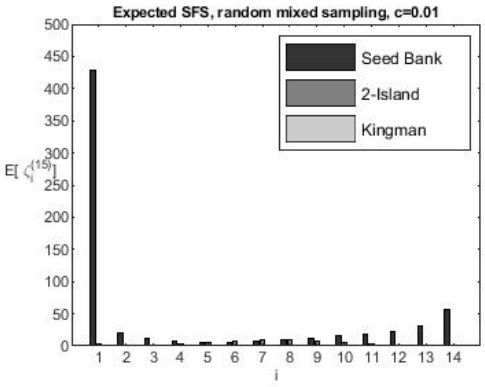


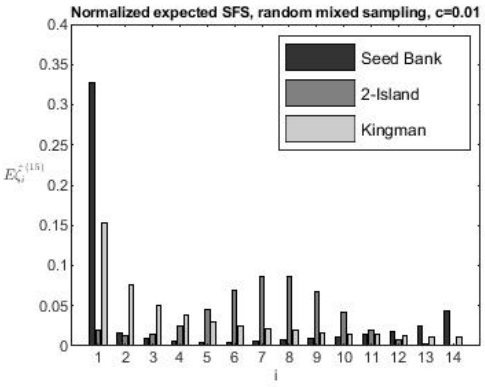

Figure 2: Expected SFS sampled from the active population, i.e. $\pi^{(15,0)}$, with $K = u = 1$.

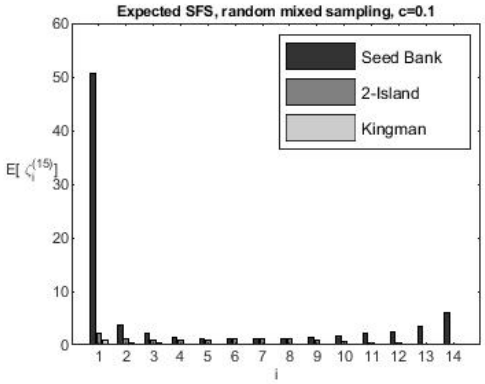
a 1 in the $i$th place, and zeros elsewhere. Then $p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)})$ solves

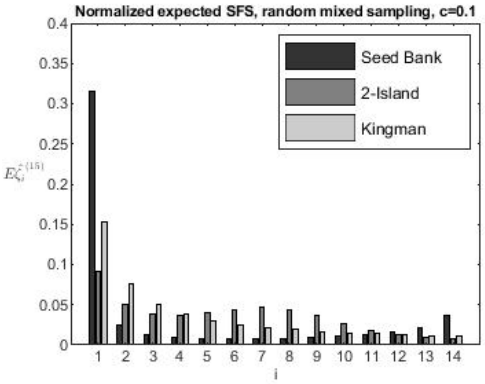
$$\begin{aligned}
&\left[ n^{(1)} \left( \frac{n^{(1)} - 1}{2} + u + c \right) + n^{(2)}(u' + Kc) \right] p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)}) \\
&= un^{(1)} \sum_{i:(n_i^{(1)}, n_i^{(2)}) = (1,0)} p(\mathbf{n}^{(1)} - \mathbf{e}_i; \mathbf{n}^{(2)}) \\
&\quad + u'n^{(2)} \sum_{i:(n_i^{(1)}, n_i^{(2)}) = (0,1)} p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)} - \mathbf{e}_i) \\
&\quad + \frac{n^{(1)}}{2} \sum_{i:n_i^{(1)} \geqslant 2} (n_i^{(1)} - 1) p(\mathbf{n}^{(1)} - \mathbf{e}_i; \mathbf{n}^{(2)}) \\
&\quad + cn^{(1)} \sum_{i:n_i^{(1)} \geqslant 1} \frac{n_i^{(2)} + 1}{n^{(2)} + 1} p(\mathbf{n}^{(1)} - \mathbf{e}_i; \mathbf{n}^{(2)} + \mathbf{e}_i) \\
&\quad + Kcn^{(2)} \sum_{i:n_i^{(2)} \geqslant 1} \frac{n_i^{(1)} + 1}{n^{(1)} + 1} p(\mathbf{n}^{(1)} + \mathbf{e}_i; \mathbf{n}^{(2)} - \mathbf{e}_i),
\end{aligned}$$

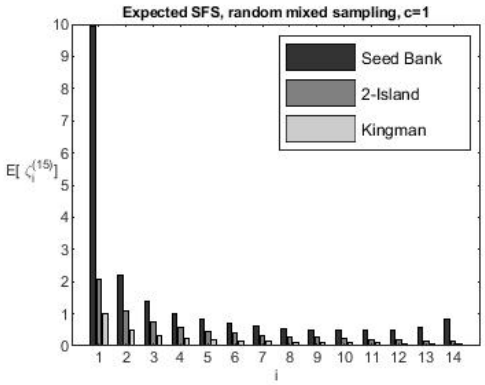
with boundary condition $p(\mathbf{e}_i; 0) = p(0; \mathbf{e}_i) = 1$. This recursion can be obtained from [DG04, Eq (2)] by omitting those transitions which are not allowed in $\mathtt{S}$, and adjusting the coefficient on the left hand side accordingly.

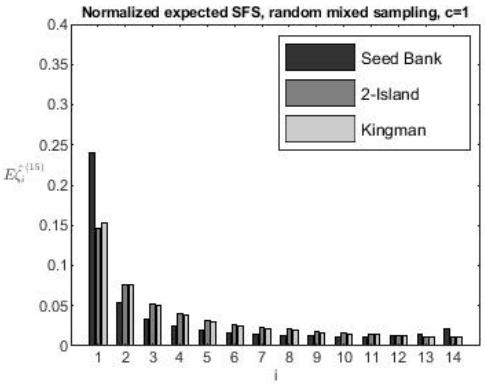

Expected SFS, random mixed sampling, c=0.01
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i

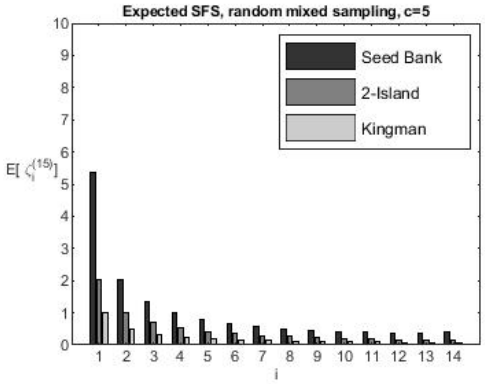

Normalized expected SFS, random mixed sampling, c=0.01
Seed Bank
2-Island
Kingman
E ζ̂_i^(15)
i

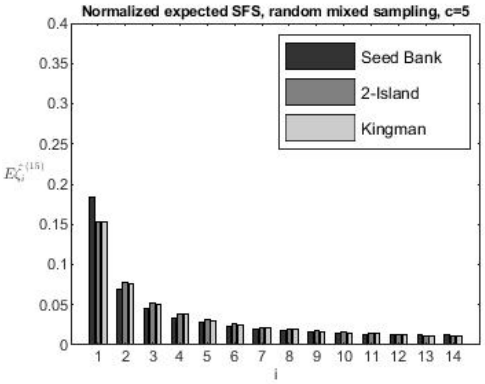

Expected SFS, random mixed sampling, c=0.1
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i


Normalized expected SFS, random mixed sampling, c=0.1
Seed Bank
2-Island
Kingman
E ζ̂_i^(15)
i


Expected SFS, random mixed sampling, c=1
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i


Normalized expected SFS, random mixed sampling, c=1
Seed Bank
2-Island
Kingman
E ζ̂_i^(15)
i


Expected SFS, random mixed sampling, c=5
Seed Bank
2-Island
Kingman
E[ ζ_i^(15) ]
i


Normalized expected SFS, random mixed sampling, c=5
Seed Bank
2-Island
Kingman
E ζ̂_i^(15)
i

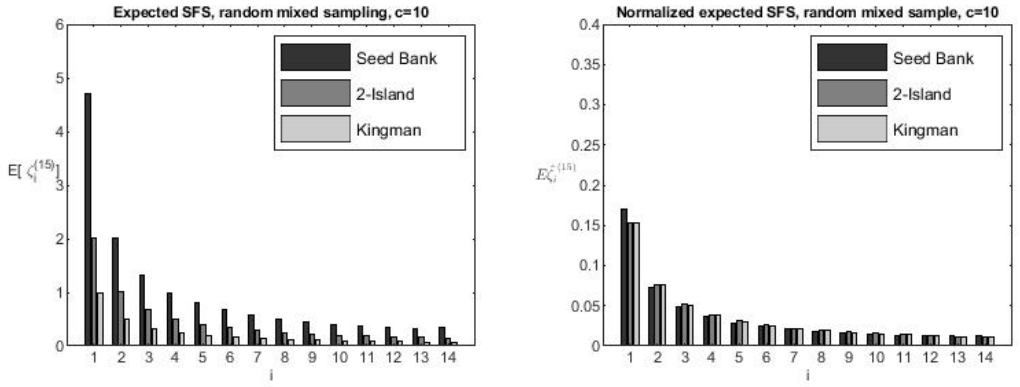



Figure 3: Expected SFS sampled from the whole population, i.e. $\pi^{\text{unif}}$, with $K = u = 1$.

### 3.2 ISM recursion

The S sampling recursion under the ISM is

$$\begin{aligned}
&\left[n^{(1)}\left(\frac{n^{(1)}-1}{2}+u+c\right)+n^{(2)}(u'+Kc)\right]p(\mathbf{t},\mathbf{n}^{(1)},\mathbf{n}^{(2)})\\
&=u\sum_{\substack{i:n_i^{(1)}=1,n_i^{(2)}=0\\ s_1^{(k)}(t_i)\neq t_j\forall j\forall k}}p(s_i^{(k)}(\mathbf{t}),\mathbf{n}^{(1)},\mathbf{n}^{(2)})+u'\sum_{\substack{i:n_i^{(1)}=0,n_i^{(2)}=1\\ s_1^{(k)}(t_i)\neq t_j\forall j\forall k}}p(s_i^{(k)}(\mathbf{t}),\mathbf{n}^{(1)},\mathbf{n}^{(2)})\\
&\quad+u\sum_{i:(n_i^{(1)},n_i^{(2)})=(1,0)}\ \sum_{(j,k):s_1^{(k)}(t_i)=t_j}(n_j^{(1)}+1)p(d_i(\mathbf{t}),d_i(\mathbf{n}^{(1)}+\mathbf{e}_j),d_i(\mathbf{n}^{(2)}))\\
&\quad+u'\sum_{i:(n_i^{(1)},n_i^{(2)})=(0,1)}\ \sum_{(j,k):s_1^{(k)}(t_i)=t_j}(n_j^{(2)}+1)p(d_i(\mathbf{t}),d_i(\mathbf{n}^{(1)}),d_i(\mathbf{n}^{(2)}+\mathbf{e}_j))\\
&\quad+n^{(1)}\sum_{i:n_i^{(1)}\geqslant 2}\frac{n_i^{(1)}-1}{2}p(\mathbf{t},\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)})\\
&\quad+cn^{(1)}\sum_{i:n_i^{(1)}\geqslant 1}\frac{n_i^{(2)}+1}{n^{(2)}+1}p(\mathbf{t},\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)}+\mathbf{e}_i)\\
&\quad+Kcn^{(2)}\sum_{i:n_i^{(2)}\geqslant 1}\frac{n_i^{(1)}+1}{n^{(1)}+1}p(\mathbf{t},\mathbf{n}^{(1)}+\mathbf{e}_i,\mathbf{n}^{(2)}-\mathbf{e}_i),
\end{aligned}$$

with boundary condition $p(\varnothing,(1),(0)) = p(\varnothing,(0),(1)) = 1$, and where $s_i^{(k)}(\mathbf{t})$ removes the $k^{\text{th}}$ element of $t_i$, e.g.

$$s_1^{(2)}((\{0,2,3\},\{1\})) = (\{0,3\},\{1\}),$$

while $d_i(\mathbf{t})$ removes $t_i$ entirely, e.g.

$$d_1((\{0,2,3\},\{1\})) = (\{1\}).$$

### 3.3 A Monte Carlo scheme for solving sampling recursions

The K and W coalescents under either IAM or parent-independent FAM are the only instances for which sampling recursions can be solved explicitly. Numerical schemes for solving the recursions directly also fail for moderate sample sizes because of combinatorial explosion of the number of equations. Hence, Monte Carlo schemes are used to approximate solutions in practice. One example of such a scheme is importance sampling, briefly introduced below.

Let $\{\mathcal{H}_g\}_{g=0}^G$ denote the history of a sample $\mathbf{n}$, so that $\mathcal{H}_0 = \mathbf{n}$, $\mathcal{H}_G$ is the type of the most recent common ancestor, and $\mathcal{H}_{g+1}$ differs from $\mathcal{H}_g$ by one coalescence, mutation, or migration event. Then the likelihood of the sample can be written as

$$\begin{aligned} p(\mathbf{n}) &= \sum_{\mathcal{H}_0,\ldots,\mathcal{H}_G} p(\mathbf{n}|\mathcal{H}_0,\ldots,\mathcal{H}_G)\mathbb{P}(\mathcal{H}_0,\ldots,\mathcal{H}_G) \\ &= \sum_{\mathcal{H}_0}\cdots\sum_{\mathcal{H}_G} p(\mathbf{n}|\mathcal{H}_0,\ldots,\mathcal{H}_G)p(\mathcal{H}_G)\prod_{g=1}^{G}\mathbb{P}(\mathcal{H}_{g-1}|\mathcal{H}_g). \end{aligned} \tag{4}$$

All of the recursions presented above are of this form, with $p(\mathbf{n}|\mathcal{H}_0,\ldots,\mathcal{H}_G) = \mathbb{1}(\mathcal{H}_0 = \mathbf{n})$, with the coefficients of the recursions denoting the transition probabilities $\mathbb{P}(\mathcal{H}_{g-1}|\mathcal{H}_g)$, and with $p(\mathcal{H}_G)$ corresponding to the boundary conditions. A naive Monte Carlo scheme for approximating this sum might sample a most recent common ancestor from the law $p(\mathcal{H}_G)$, evolve the sample stochastically until it reaches the desired size $n+1$ with probabilities given by the coefficients of the appropriate sampling recursion, and then evaluate the quantity of interest $\mathbb{1}(\mathcal{H}_0 = \mathbf{n})$, where $\mathcal{H}_0$ is the last sample with size $n$. However, likelihoods in genetics can be vanishingly small, which renders the number of such simulations required for accurate estimators infeasibly large. Instead, we introduce an importance sampling proposal distribution $\mathbb{Q}(\mathcal{H}_g|\mathcal{H}_{g-1})$, which acts in the opposite direction of time to $\mathbb{P}(\mathcal{H}_{g-1}|\mathcal{H}_g)$, i.e. from the observed leaves towards the most recent common ancestor, and rewrite the summation in (4) as

$$p(\mathbf{n}) = \sum_{\mathcal{H}_0}\cdots\sum_{\mathcal{H}_G} p(\mathcal{H}_G)\prod_{g=1}^{G}\frac{\mathbb{P}(\mathcal{H}_{g-1}|\mathcal{H}_g)}{\mathbb{Q}(\mathcal{H}_g|\mathcal{H}_{g-1})}\mathbb{Q}(\mathcal{H}_g|\mathcal{H}_{g-1}).$$

We will specify $\mathbb{Q}$ in such a way that $\mathbb{Q}(\mathcal{H}_0 = \mathbf{n}) = 1$, which is why the factor $p(\mathbf{n}|\mathcal{H}_0,\ldots,\mathcal{H}_G)$ no longer appears. This initial condition is then propagated back to the most recent common ancestor with yet-to-be-specified transition probabilities $\mathbb{Q}(\mathcal{H}_g|\mathcal{H}_{g-1})$, and once the most recent common ancestor is reached, we evaluate the modified quantity of interest

$$p(\mathcal{H}_G)\prod_{g=1}^{G}\frac{\mathbb{P}(\mathcal{H}_{g-1}|\mathcal{H}_g)}{\mathbb{Q}(\mathcal{H}_g|\mathcal{H}_{g-1})}.$$

Every sample results in a positive contribution under this scheme, reducing the variance of estimators. Careful choices of $\mathbb{Q}$ can reduce variance even further.

The zero-variance proposal distribution $\mathbb{Q}$ under K (and thus also W) was described in [SD00], and extended to TI in [DG04]. Neither can be implemented, but both articles

also provide heuristic approximations from which ancestral coalescence, mutation, and migration events can be sampled, and which result in low variance estimators in practice. In this section we present similar heuristics for `S` under the IAM and ISM. As before, corresponding results under the FAM are provided in the appendix.

For the IAM and ISM, we suggest the following procedure for sampling the next event backwards in time given that the current state is $(\mathbf{n}^{(1)}, \mathbf{n}^{(2)})$:

(i) Sample the active or dormant subpopulation with probabilities proportional to

$$\left( n^{(1)} \left( \frac{n^{(1)} - 1}{2} + c + u \right), n^{(2)}(Kc + u') \right).$$

Denote the chosen subpopulation by $j$.

(ii) Sample a lineage uniformly at random from subpopulation $j$. Denote its allele by $i$.

(iii) With probabilities proportional to

$$\left( \frac{(n_i^{(j)} - 1)^+}{2} \mathbb{1}_{\{j=1\}}, u\mathbb{1}_{\{j=1\}} + u'\mathbb{1}_{\{j=2\}}, c\mathbb{1}_{\{j=1\}} + Kc\mathbb{1}\{j = 2\} \right),$$

merge the lineage with another one with allele $i$ on island $j$, remove from type $i$ a randomly chosen mutation that does not appear on any other lineage, or migrate the lineage to the other subpopulation. The mutation probability is taken to be 0 if there are no eligible mutations on the lineage, or if the frequency of the allele is greater than one in the case of the IAM. For the IAM, we also interpret the removal of a mutation as the removal of the lineage from the sample.

# 4 Inference and model selection

In this section we provide an example of the impact of the presence or absence of a seed bank on model selection, and on estimating coalescent parameters from genetic data. Our focus is on the full likelihood Monte Carlo methods introduced in Section 3.3, rather than on summary statistics such as the (n)SFS. While computationally intensive, this choice lets us draw robust conclusions about the extent to which DNA sequence data can distinguish between our three model classes even in principle, without further confounding by the limitations of any particular summary statistic.

## 4.1 Model selection based on sampling formulas

We used a pseudo-marginal Metropolis-Hastings algorithm [AR09] to perform model selection and parameter inference simultaneously for models `K`, `S`, and `TI` using the full likelihood of the observed sequence data. Model `W` was not included as it is not identifiable from `K`. We focus on the ISM in order to balance biological relevance and computational cost. Data set of 100 observed, non-recombining sequences with were

simulated under each model and various parameter regimes to act as observed data. In each case, all 100 sequences were sampled from island 1 to model the impact of an unknown seed bank or population subdivision.

The state space of our pseudo-marginal Markov chain consists of the model indicator $\mathtt{I} \in \{\mathtt{K}, \mathtt{S}, \mathtt{TI}\}$, as well as seven non-negative variables

$$\Theta := (u_{\mathtt{K}}, u_{\mathtt{S}}, u_{\mathtt{TI}}, c_{\mathtt{S}}, c_{\mathtt{TI}}, K_{\mathtt{S}}, K_{\mathtt{TI}}).$$

In particular, the fact that $u' = 0$ under $\mathtt{S}$ and $\mathtt{TI}$ was assumed to be known. Given an observed data set $(\mathbf{t}, \mathbf{n})$, the target distribution is the posterior

$$q(\mathtt{I}, \Theta | \mathbf{t}, \mathbf{n}) \propto p(\mathbf{t}, \mathbf{n} | \mathtt{I}, \Theta) q_{\mathtt{I}}(\mathtt{I}) q_{u_{\mathtt{K}}}(u_{\mathtt{K}}) \prod_{J \in \{\mathtt{S}, \mathtt{TI}\}} q_{u_J}(u_J) q_{c_J}(c_J) q_{K_J}(K_J),$$

where $\mathbf{n} = (\mathbf{n}^{(1)}, \mathbf{n}^{(2)})$ in the case of scenarios $\mathtt{S}$ and $\mathtt{TI}$. Here, the likelihood $p(\mathbf{t}, \mathbf{n} | \mathtt{I}, \Theta)$ only charges those coordinates of $\Theta$ that play a role for model $\mathtt{I}$, and is flat in all other directions. The prior distributions are $q_{\mathtt{I}} = (1/3, 1/3, 1/3)$, and Gamma-distributions with shape parameter 4 for all other variables. Scale parameters are fixed at $1/4$ for the $c$ and $K$-variables, and by requiring the prior mean to equal the corresponding Watterson estimator for the $u$-variables. This updating of locally redundant variables increases model dimension, but also results in faster mixing across the three different models since all parameters are updated simultaneously (see the "saturated space approach" of [BGR09]), and accounts for the fact that the same number of segregating sites can fit two very different mutation rates in different model classes.

The model index was resampled uniformly at random at each time step, including the possibility of remaining in place. All other parameters were updated using independent Gaussian increments with mean 0 and variance $\approx 1/14$, with all parameters reflected at zero. The importance sampling scheme of Section 3.3 was used to obtain unbiased estimates of likelihoods, with particle numbers set to 400 for $\mathtt{K}$, and 20 000 for $\mathtt{S}$ and $\mathtt{TI}$. Variances of estimators were further reduced by employing stopping time resampling [Jen12]. These parameters were calibrated so that the log-likelihood estimator variances were close to 3, and acceptance probabilities close to 7%, shown to be optimal in [STRR15]. `C++` code for both simulating observed data sets, and conducting the inference described above, is available at `https://github.com/JereKoskela/seedbank-infer`.

Three realizations of this Markov chain, one for each simulated data set, were run for 100 000 steps each, initialized from a uniformly chosen model, and the continuous parameters initialized from their respective prior means. The most immediate question is whether each data-generating model can be correctly recovered from its observed data set. Table 1 provides marginal posterior probabilities of each model and data set. It is evident that the true model can be recovered from a moderate amount of data with high confidence for moderate second population sizes and migration rates. However, as might be expected, large migration rate or small second populations make it challenging to tell the three models apart, at least when only sampling a single locus from one population.

Posterior distributions of parameters given a model class are also of interest. These are summarized in Table 2. None of the parameters are strongly identified, and in

| True model | $q_{\mathtt{I}}(\mathtt{K}\|\mathbf{t},\mathbf{n})$ | $q_{\mathtt{I}}(\mathtt{S}\|\mathbf{t},\mathbf{n})$ | $q_{\mathtt{I}}(\mathtt{TI}\|\mathbf{t},\mathbf{n})$ | $u_0$ | $c_0$ | $K_0$ |
|---|---|---|---|---|---|---|
| K | 0.950 | 0.042 | 0.008 | 10 | - | - |
| S | 0.000 | 1.000 | 0.000 | 10 | 1 | 1 |
| TI | 0.132 | 0.027 | 0.841 | 10 | 1 | 1 |
| S | 0.258 | 0.463 | 0.279 | 10 | 5 | 1 |
| TI | 0.439 | 0.028 | 0.533 | 10 | 5 | 1 |
| S | 0.224 | 0.519 | 0.257 | 10 | 1 | 5 |
| TI | 0.356 | 0.137 | 0.507 | 10 | 1 | 5 |

Table 1: Marginal posterior probabilities of each model class for given data-generating parameters.

some cases estimates are biased due to the fact that our data set consists of only a single locus. The mutation rate was the slowest to mix in all cases (results not shown).

| | 0.025 | 0.5 | 0.975 | Model | 0.025 | 0.5 | 0.975 | Model |
|---|---|---|---|---|---|---|---|---|
| $u$ | 6.30 | 8.97 | 11.3 | K | | | | |
| $u$ | 7.45 | 8.60 | 9.87 | S | 13.0 | 16.3 | 19.7 | TI |
| $c$ | 0.20 | 0.98 | 1.86 | S | 0.39 | 1.02 | 1.94 | TI |
| $K$ | 0.09 | 0.54 | 1.72 | S | 0.25 | 0.84 | 1.82 | TI |
| $u$ | 8.32 | 8.59 | 9.06 | S, $c = 5$ | 6.26 | 9.69 | 12.9 | TI, $c = 5$ |
| $c$ | 2.00 | 3.44 | 4.23 | S, $c = 5$ | 2.04 | 5.06 | 9.02 | TI, $c = 5$ |
| $K$ | 0.57 | 1.04 | 1.37 | S, $c = 5$ | 0.35 | 0.86 | 1.87 | TI, $c = 5$ |
| $u$ | 4.06 | 6.03 | 8.97 | S, $K = 5$ | 8.29 | 13.1 | 16.8 | TI, $K = 5$ |
| $c$ | 0.36 | 1.03 | 2.52 | S, $K = 5$ | 0.28 | 1.04 | 2.39 | TI, $K = 5$ |
| $K$ | 1.69 | 4.23 | 6.91 | S, $K = 5$ | 1.68 | 3.47 | 7.01 | TI, $K = 5$ |

Table 2: Posterior quantiles for various parameters and scenarios. Where not specified, the parameters are $u = 10$, $c = K = 1$. All estimates are conditional on the true model class.

Low migration rates and large second populations are also problematic, albeit in a different way. Figure 4 shows the empirical distribution of the number of segregating sites for samples drawn from models S and TI. Results with $c = 0.2$ or $K = 0.2$ have noticeably broader supports and heavier tails than any of the other scenarios, because a migration from one population to another is rare, but will result in a very long ancestral tree if it takes place. The consequence for inference is that realisations of data sets are not informative of the model or parameters which generated them, and the importance sampling schemes from Section 3.3 will also suffer from high variance and very long run times. This is reminiscent of Figures 2 and 3, which show that the expected SFS detects an excess of singletons due to a strong seed bank under uniform sampling, but not when only the active population is sampled.

While the method presented in this section does not scale to large data sets, it sets a benchmark for what can be expected of the performance of more scalable methods. In particular, the three model classes ought to be distinguishable with high confidence even in the presence of parameter uncertainty, provided that the true migration rates

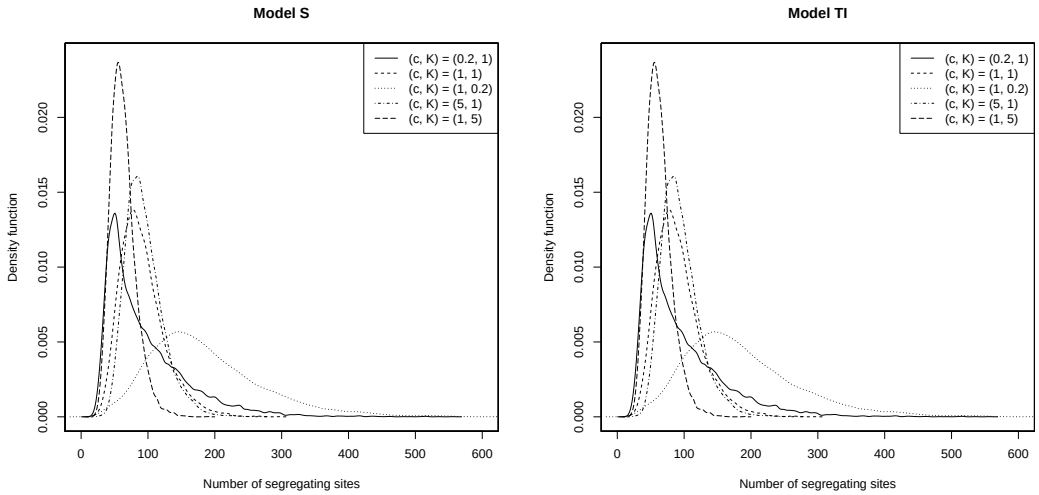



Figure 4: Empirical distributions of the number of segregating site among 100 samples from population 1, estimated from 10 000 simulations.

and seed bank sizes are moderate. True values of $c$ or $K$ that are too large are problematic due to less statistical separation between the models, while low data-generating values of $c$ and $K$ cause instability both in observed data and in our Monte Carlo scheme. Estimating precise values of parameters within model classes is challenging without strong prior information, or data from multiple unlinked loci, neither of which has been used in our model selection pipeline.

## 4.2 Detecting mutation in the seed bank

In this section we focus on a different model selection problem: whether mutation is taking place in a strong seed bank that is known to be present. Data sets were simulated under two scenarios with a moderate seed bank and migration rate $K = c = 1$:

S1. Model `S` with $u = 10, u' = 0$.

S2. Model `S` with $u = u' = 5$.

All other simulation details are as in Section 4.1. A pseudo-marginal Metropolis-Hastings chain was run targeting these two hypotheses, with the same priors as in Section 4.1. In scenario S1 we assumed that $u' = 0$ was known, while in scenario S2 we assumed that $u = u'$ was known, but that the common value itself was not. The posterior probabilities of each scenario are given in Table 3.

It is evident that the presence or absence of mutation in a seed bank can be detected with high confidence from a modest amount of data. Table 4 shows that parameters remain only weakly identified and that estimates can be biased, particularly in the case of mutation rates. Once again, mutation rates were also the slowest parameters to mix.

| True scenario | $q_I(S1\|\mathbf{t},\mathbf{n})$ | $q_I(S2\|\mathbf{t},\mathbf{n})$ |
|---|---|---|
| `S1` | 1.000 | 0.000 |
| `S2` | 0.098 | 0.902 |

Table 3: Marginal posterior probabilities of each scenario.

| | 0.025 | 0.5 | 0.975 | Scenario |
|---|---|---|---|---|
| $u$ | 3.42 | 5.82 | 9.02 | S1 |
| $c$ | 0.17 | 0.98 | 2.89 | S1 |
| $K$ | 0.15 | 0.80 | 1.98 | S1 |
| $u = u'$ | 2.29 | 5.54 | 8.57 | S2 |
| $c$ | 0.24 | 1.04 | 2.23 | S2 |
| $K$ | 0.28 | 0.92 | 2.04 | S2 |

Table 4: Posterior quantiles for various parameters and scenarios. The columns labeled $(u_0, c_0, K_0)$ denote the corresponding data-generating parameters.

# 5 Discussion

We have reviewed several population genetic models related to seed banks, in combination with several classical mutation models. We derived expressions for classical population genetic summary statistics such as the $F_{ST}$ and the SFS for various combinations of coalescent and mutation models. We then studied the identifiability of various scenarios and parameters based on tractable summary statistics, as well as computationally intensive full likelihood methods. Throughout, our focus is on deriving and testing generic methodology without prior assumptions on mutation models, parameter ranges, or model classes. This is to facilitate analysis of sequence data across a wide range of scenarios and species.

Explicit expressions for $F_{ST}$ for the IAM and ISM can be obtained using phase-type distribution arguments [HSJB19]. A strong seed bank produces elevated levels of $F_{ST}$, but less so than the two-island model with identical parameters. The signal is slightly stronger in the case without mutation in the seed bank compared to the case with mutation, but generally appears to be too weak to allow for confident detection of a strong seed bank. The FAM also yields similar results (see appendix).

Considering the normalized SFS instead of $F_{ST}$ results in improved statistical power. The Kingman and the weak seed bank scenarios can only be distinguished with prior knowledge of the germination rate or population-rescaled mutation rate(s), whereupon the number of expected segregating sites suffices as a statistic. The strong seed bank and two island models result in an excess of singletons and a lighter tail in the nSFS when compared to the classical Kingman case, for sample sizes as low as $n = 15$. Thus, these two scenarios can be distinguished from `K` and `W`, but not from each other. The deviation of the nSFS from the Kingman coalescent arises due to the reduced (or vanishing) mutation rate in the seed bank, and so sampling dormant lineages is an effective way to boost the power of the model selection procedure.

To study the scope of possible inference, we used a Monte Carlo scheme to approximate likelihoods of observed sequence data. For moderate data-generating parameter

values, model selection from simulated data gave good results for samples of size $n = 100$ and a single locus, even in the presence of parameter uncertainty. Accounting for parameter uncertainty in the simulation pipeline is particularly important, because standard estimators such as the Watterson estimator assume a fixed coalescent model, and thus using the wrong estimator can strongly bias further inferences. Frequent migration or a small relative second sub-population size cause diminished statistical power, while rare migrations or a large second sub-population cause instabilities in both the model and in standard importance sampling estimators of likelihoods.

We have also demonstrated that our method is able to detect whether mutation is taking place in the seed bank, again in the presence of parameter uncertainty. Thus, it provides a promising first step towards answering similar questions in general [LJ11], such as assessing molecular clock hypotheses for bacteria, or other organisms without easy access to a fossil record [Mau07].

Knowledge of the real substitution rate $\hat{\mu}$ per year at the (active) locus under consideration would allow a real-time embedding of the coalescent history via

$$\text{coalescent time unit } \times u_{\mathtt{I}} \approx \text{ year } \times \mu,$$

for $\mathtt{I} \in \{\mathtt{K}, \mathtt{W}, \mathtt{S}, \mathtt{TI}\}$ (see e.g. [EBBF15, Eq (4)], [SBB13, Section 4.2]). This allows the estimation of quantities such as the TMRCA of a sample in real time, not only in units of coalescent time. Typically, one coalescent time unit corresponds to $O(N)$ generations under all four models considered in this paper.

Our paper is a starting point for the statistical methodology for seed bank detection. We have shown that model selection and inference are possible from moderate data sets in principle, but several important points remain to be addressed.

First, the adequacy and universality of the models needs to be established. They all describe idealized scenarios in population genetics, with constant population sizes, and in the absence of further evolutionary forces such as selection or demography. The effect of such forces in the presence of seed banks remains unknown, and may confound some or all of the results we have presented. Indeed [ŽT12, SAMT19] have shown that weak seed banks and demographic changes confound each other unless considered jointly, and similar results are available for the structured coalescent [MRG$^+$16, RMG$^+$18]. A similar analysis for the strong seed bank model, and the effect of a misspecified model on demographic inference, remains an open problem. A layer of complexity is added in the modeling of demography with a seed bank through the effect of changes in demography for the dormant population. In order to obtain a (time-changed) coalescent with a strong seed bank mechanism, the demographic changes would have to equally affect the active and the dormant population. If, however, the seed bank remains constant relative to the demographic changes in the active population, the result would be a coalescent with a time-inhomogeneous (de)activation mechanism.

Second, the type of seed bank formation mechanism itself needs to be discussed. The strong seed bank model of [BGKW16] analyzed here follows the modeling idea of "spontaneous switching" in [LJ11], where switching between the active and the dormant state happens on an individuals basis. [LJ11] argue that this model might be

appropriate for populations in "stable" environments, but that in real populations initiation of or resuscitation from dormancy can be triggered by environmental cues, leading to "responsive switching" where many individuals switch their state simultaneously. This mechanism can be incorporated at the same scale as the spontaneous switching and leads to a scaling limit that is different from the migration-type behavior of the strong seed bank model (and of course also differs from the weak seed bank model), cf. [BGKW19], as it includes simultaneous activation and deactivation of lineages. The effect of this additional mechanism on the statistics discussed here remains to be studied.


## Acknowledgements

JB was supported by DFG Priority Programme 1590 "Probabilistic Structures in Evolution", project 1105/5-1. EB was supported by DFG RTG 1845 and BMS Berlin Mathematical School. JK was supported in part by EPSRC grant EP/R044732/1.


## Appendix

### Classical measures of population structure under the FAM

The *sample heterozygosity* $H$ of a population is defined as the probability of two individuals drawn independently and uniformly from the population carrying different alleles. For $\mathtt{K}$ and $\mathtt{W}$, the stationary sample heterozygosity is

$$H^{\mathtt{K}} := 2\mathbb{E}^{\mathtt{K}}[X(1-X)], \qquad \text{and} \qquad H^{\mathtt{W}} := 2\mathbb{E}^{\mathtt{W}}[X(1-X)],$$

where $X$ has the stationary distribution of (1) corresponding to each model.

A well-known result (e.g. [Eth11, p. 49]) states that

$$H^{\mathtt{K}} = \frac{4u_1u_2}{(u_1+u_2)(1+2u_1+2u_2)},$$

an similarly we have the intuitive result

$$H^{\mathtt{W}} = \frac{4u_1u_2}{(u_1+u_2)(\beta^2+2u_1+2u_2)}.$$

For structured populations one distinguishes between the *global* and *local* sample heterozygosities, corresponding to samples taken from the overall population, resp. from each sub-population. Thus, with $(X,Y)$ being the solution to (1) at stationarity, the local sample heterozygosities for each sub-population under $\mathtt{S}$ and $\mathtt{TI}$ are

$$H_X^{\mathtt{S}} := 2\mathbb{E}^{\mathtt{S}}[X(1-X)], \qquad H_X^{\mathtt{TI}} := 2\mathbb{E}^{\mathtt{TI}}[X(1-X)],$$
$$H_Y^{\mathtt{S}} := 2\mathbb{E}^{\mathtt{S}}[Y(1-Y)], \qquad H_Y^{\mathtt{TI}} := 2\mathbb{E}^{\mathtt{TI}}[Y(1-Y)],$$

and therefore the global sample heterozygosities can be written as

$$\begin{aligned} H^{\mathtt{S}} &:= \frac{K^2}{(K+1)^2} H_X^{\mathtt{S}} + \frac{2K}{(K+1)^2} \mathbb{E}_{\mu^{\mathtt{S}}}[X(1-Y)+Y(1-X)] + \frac{1}{(K+1)^2} H_Y^{\mathtt{S}}, \\ H^{\mathtt{TI}} &:= \frac{K^2}{(K+1)^2} H_X^{\mathtt{TI}} + \frac{2K}{(K+1)^2} \mathbb{E}_{\mu^{\mathtt{TI}}}[X(1-Y)+Y(1-X)] \\ &\quad + \frac{1}{(K+1)^2} H_Y^{\mathtt{TI}}, \end{aligned} \tag{5}$$

where the weights on the local heterozygosities are the probabilities associated with sampling two lineages uniformly at random from the global population. The sample heterozygosity at stationarity is well-studied under the FAM and either `K` or `TI` [Her94], it has so far not been considered for seed banks.

Note that we can rewrite the sample heterozygosities for $\mathtt{I} \in \{\mathtt{S}, \mathtt{TI}\}$ in terms of mixed moments using the notation

$$M_{n,m}^{\mathtt{I}} := \mathbb{E}_{\mu^{\mathtt{I}}}[X^n Y^m], \quad n, m \geqslant 0.$$

This immediately gives

$$H_X^{\mathtt{I}} = 2(M_{1,0}^{\mathtt{I}} - M_{2,0}^{\mathtt{I}}), \qquad H_Y^{\mathtt{I}} = 2(M_{0,1}^{\mathtt{I}} - M_{0,2}^{\mathtt{I}}),$$

and therefore

$$H^{\mathtt{I}} = \frac{2}{(K+1)^2}\Big((K^2+K)M_{1,0}^{\mathtt{I}} + (K+1)M_{0,1}^{\mathtt{I}} - 2KM_{1,1}^{\mathtt{I}} - K^2 M_{2,0}^{\mathtt{I}} - M_{0,2}^{\mathtt{I}}\Big).$$

These mixed moments can be calculated recursively [BBGW19, Lemma 2.7]. For example, $M_{0,0}^{\mathtt{I}} = 1$ and

$$\begin{aligned} M_{1,0}^{\mathtt{I}} &= \frac{cu_2' + u_1u_2' + u_2u_2' + cKu_2}{cu_1' + cu_2' + u_1u_1' + u_1u_2' + u_2u_1' + u_2u_2' + cKu_1 + cKu_2}, \\ M_{0,1}^{\mathtt{I}} &= \frac{cu_2' + u_1'u_2 + u_2u_2' + cKu_2}{cu_1' + cu_2' + u_1u_1' + u_1u_2' + u_2u_1' + u_2u_2' + cKu_1 + cKu_2}, \end{aligned}$$

for the first moments. These first moments do not depend on $\alpha$ and $\alpha'$, which is clear intuitively since they represent variance parameters. Hence, $M_{1,0}^{\mathtt{I}}$ and $M_{0,1}^{\mathtt{I}}$ are invariant for $\mathtt{I} \in \{\mathtt{TI}, \mathtt{S}\}$. The expression for the second moments can also be computed easily, but are cumbersome and therefore omitted.

In the case of equal relative population sizes ($K = 1$), migration rate $c = 1$ and mutation rates $u_1 = u_2 = u_1' = u_2' = 1/2$, we obtain

$$H^{\mathtt{S}} = \frac{14}{31} \approx 0.4516 \; > H^{\mathtt{TI}} = \frac{13}{32} \approx 0.4063 \; > \; \frac{1}{3} = H^{\mathtt{K}}.$$

Moreover, using simple sign arguments, we find that these relationships also hold in a more general context: if $u_1 = u_1'$, $u_2 = u_2'$, and $K = 1$, then for all $u_1, u_2, c \geqslant 0$ we have $H^{\mathtt{S}} \geqslant H^{\mathtt{TI}} \geqslant H^{\mathtt{K}}$. However, in all other cases (e.g. $c = u_1 = u_2 = u_1' = u_2' = 1$, $K = 0.01$), the second inequality does not hold. It is also interesting to note that if

$\beta^2 < 3/14$, then $H^{\mathtt{W}} = 1/(\beta^2+2) > H^{\mathtt{S}}$, showing that a weak seed bank can generate more heterozygosity than a strong seed bank in some cases.

Overall, scenario S has elevated levels of genetic variability relative to TI or K at stationarity. The TI sample heterozygosity is somewhat lower, which is consistent with the idea that genetic drift in the second island reduces variability.

**Remark 5.1.** If we naively let $K \to \infty$ (i.e. the relative second island size $\to 0$) in equation 5, ignoring the intrinsic dependence of the variables $X$ and $Y$ on this parameter, we recover the sample heterozygosity of K,

$$H_X^{\mathtt{S}} \to H^{\mathtt{K}} \qquad \text{and} \qquad H_X^{\mathtt{TI}} \to H^{\mathtt{K}}.$$

This convergence holds in a stronger sense on the diffusion level, and will be discussed theoretically in related future work.

**Remark 5.2.** The stationary sample heterozygosity cannot distinguish between K and W if neither the germination rate $\beta$ nor the population-rescaled mutation rate $u$ are known. But K and W can be differentiated using, for example, the *rate of decay* of sample heterozygosity over time *in the absence of mutation*. Define

$$H^{\mathtt{I}}(t,x) := 2\mathbb{E}^{\mathtt{I}}[X(t)(1-X(t))|X(0)=x],$$

for $I \in \{\mathtt{K},\mathtt{W}\}$. Then we obtain

$$H^{\mathtt{K}}(t,x) = 2e^{-t}x(1-x), \qquad \text{while} \qquad H^{\mathtt{W}}(t,x) = 2e^{-\beta^2 t}x(1-x),$$

showing that $H^{\mathtt{W}}$ decays more slowly than $H^{\mathtt{K}}$ due to the seed bank slowing down genetic drift [KKL01].

**Wright's $F_{ST}$ for the FAM** In the previous section we derived the sample heterozygosities, i.e. the probabilities of sampling *distinct* types, in the FAM. The probabilities of sampling *identical* types are simply their complements, yielding

$$F_{ST}^{\mathtt{I}} = \frac{(K+1)H^{\mathtt{I}} - KH_X^{\mathtt{I}} - H_Y^I}{(K+1)H^{\mathtt{I}}}$$

for $\mathtt{I} \in \{\mathtt{S},\mathtt{TI}\}$. For example, fixing $u_1 = u_2 = 1/2 = u_1' = u_2', c = K = 1$ and $\alpha = 1$, TI ($\alpha' = 1$) leads to a stronger differentiation than S ($\alpha' = 0$),

$$F_{ST}^{\mathtt{S}} = \frac{1}{28} < \frac{1}{13} = F_{ST}^{\mathtt{TI}},$$

again indicating that strong seed banks introduce some population substructure, but that the effect is stronger in the two island model. This is intuitive, because the dynamics of the population are closer to those of two independent subpopulations when both demes undergo genetic drift than when only one subpopulation does.

Figure 5 further illustrates how $F_{ST}$ depends on the model parameters in both cases. The first plot shows $F_{ST}$ as a function of the migration rate $c$. As expected, $F_{ST}$ approaches 0 as $c$ increases, leading to a well-mixed population, and the $F_{ST}$ of

TI dominates the one of S by a factor of approximately 2.1 for these parameters. The second plot shows $F_{ST}$ as a function of the mutation rate, with similar results. This is again in accordance with expectation, since increasing mutation rates in both subpopulations further mixes the population. The third plot shows the dependence of $F_{ST}$ on the relative population size $K$. The $F_{ST}$ is nearly 0 if the relative population size on either island is very small (i.e. $K$ very small or very large), as this results in a small probability of sampling two individuals from different demes when sampling uniformly from the whole population.

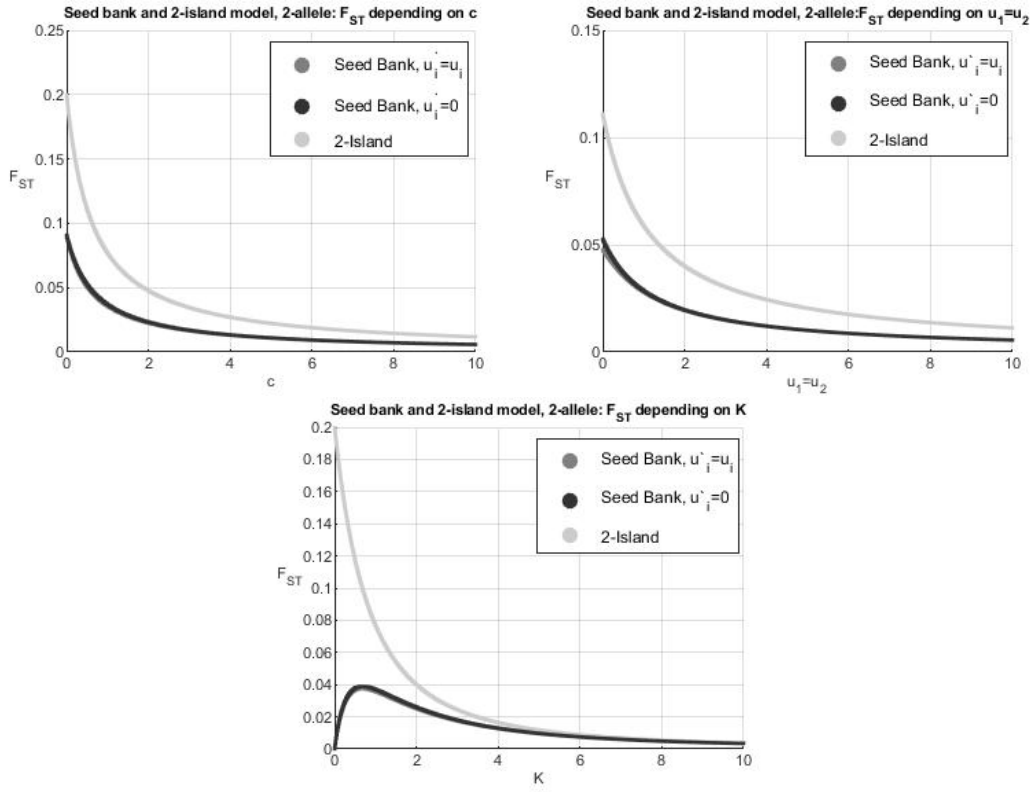



Figure 5: $F_{ST}$ under S and TI as a function of various parameters in the FAM. Where not specified, $K = c = 1$, $u_1 = u_2 = 0.5$.

In the absence of mutation in the seed bank, $u' = 0$, and with the parameters $u_1 = u_2 = 1/2, K = c = 1$, we get

$$F_{ST}^{\mathtt{S}} = \frac{1}{27} > \frac{1}{28},$$

a *slightly* stronger signal than in the case with mutation. The relationship between $K$, $c$ and the $F_{ST}$ in this setting is also illustrated in Figure 5.

## Recursions for the FAM sampling distribution

Under S and the FAM, the sampling distribution $p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)})$ solves

$$\begin{aligned}
&\left[ n^{(1)} \left( \frac{n^{(1)} - 1}{2} + u_1 + u_2 + c \right) + n^{(2)}(u_1' + u_2' + Kc) \right] p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)}) \\
&= u_2(n_1^{(1)} + 1)\mathbb{1}(n_2^{(1)} > 0) p(\mathbf{n}^{(1)} + \mathbf{e}_1 - \mathbf{e}_2; \mathbf{n}^{(2)}) \\
&\quad + u_1(n_2^{(1)} + 1)\mathbb{1}(n_1^{(1)} > 0) p(\mathbf{n}^{(1)} - \mathbf{e}_1 + \mathbf{e}_2; \mathbf{n}^{(2)}) \\
&\quad + u_2'(n_1^{(2)} + 1)\mathbb{1}(n_2^{(2)} > 0) p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)} + \mathbf{e}_1 - \mathbf{e}_2) \\
&\quad + u_1'(n_2^{(2)} + 1)\mathbb{1}(n_1^{(2)} > 0) p(\mathbf{n}^{(1)}; \mathbf{n}^{(2)} - \mathbf{e}_1 + \mathbf{e}_2) \\
&\quad + n^{(1)} \frac{n_1^{(1)} - 1}{2} p(\mathbf{n}^{(1)} - \mathbf{e}_1; \mathbf{n}^{(2)}) + n^{(1)} \frac{n_2^{(1)} - 1}{2} p(\mathbf{n}^{(1)} - \mathbf{e}_2; \mathbf{n}^{(2)}) \\
&\quad + cn^{(1)} \frac{n_1^{(2)} + 1}{n^{(2)} + 1} \mathbb{1}(n_1^{(1)} > 0) p(\mathbf{n}^{(1)} - \mathbf{e}_1; \mathbf{n}^{(2)} + \mathbf{e}_1) \\
&\quad + cn^{(1)} \frac{n_2^{(2)} + 1}{n^{(2)} + 1} \mathbb{1}(n_2^{(1)} > 0) p(\mathbf{n}^{(1)} - \mathbf{e}_2; \mathbf{n}^{(2)} + \mathbf{e}_2) \\
&\quad + Kcn^{(2)} \frac{n_1^{(1)} + 1}{n^{(1)} + 1} \mathbb{1}(n_1^{(2)} > 0) p(\mathbf{n}^{(1)} + \mathbf{e}_1; \mathbf{n}^{(2)} - \mathbf{e}_1) \\
&\quad + Kcn^{(2)} \frac{n_2^{(1)} + 1}{n^{(1)} + 1} \mathbb{1}(n_2^{(2)} > 0) p(\mathbf{n}^{(1)} + \mathbf{e}_2; \mathbf{n}^{(2)} - \mathbf{e}_2),
\end{aligned}$$

where $\mathbb{1}(E) = 1$ if event $E$ is true, and 0 otherwise. Boundary conditions are typically prescribed as the stationary distribution specified by the mutation rates, at least when $u_1 = u_1'$ and $u_2 = u_2'$:

$$\begin{aligned}
p((1,0);(0,0)) = p((0,0);(1,0)) = \rho_1, \\
p((0,1);(0,0)) = p((0,0);(0,1)) = \rho_2.
\end{aligned}$$

## A Monte Carlo scheme for the FAM recursions

Let $p_i(\mathbf{e}_j | \mathbf{n}^{(1)}, \mathbf{n}^{(2)})$ denote the probability that a further lineage sampled from island $i \in \{1,2\}$ carries allele $j \in \{1,2\}$, given observed allele frequencies $\mathbf{n}^{(1)}, \mathbf{n}^{(2)}$ from islands 1 and 2, respectively. These conditional sampling distributions are intractable, but as outlined in Section 3.3, approximating them will produce efficient algorithms. Let

$$D(n^{(1)}, n^{(2)}) := n^{(1)} \left( \frac{n^{(1)} - 1}{2} + u + c \right) + n^{(2)}(u' + Kc).$$

A calculation similar to [SD00, Theorem 1] identifies the zero-variance proposal dis-

tribution for the FAM as

$$(\mathbf{n}^{(1)},\mathbf{n}^{(2)}) \mapsto (\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)}) \text{ w. prob. } \frac{n_i^{(1)}(n_i^{(1)}-1)/2}{p_1(\mathbf{e}_i|\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)})D(n^{(1)},n^{(2)})},$$

$$(\mathbf{n}^{(1)},\mathbf{n}^{(2)}) \mapsto (\mathbf{n}^{(1)}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}^{(2)}) \text{ w. prob. } \frac{un_i^{(1)}p_1(\mathbf{e}_j|\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)})}{p_1(\mathbf{e}_i|\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)})D(n^{(1)},n^{(2)})},$$

$$(\mathbf{n}^{(1)},\mathbf{n}^{(2)}) \mapsto (\mathbf{n}^{(1)},\mathbf{n}^{(2)}-\mathbf{e}_i+\mathbf{e}_j) \text{ w. prob. } \frac{u'n_i^{(2)}p_2(\mathbf{e}_j|\mathbf{n}^{(1)},\mathbf{n}^{(2)}-\mathbf{e}_i)}{p_2(\mathbf{e}_i|\mathbf{n}^{(1)},\mathbf{n}^{(2)}-\mathbf{e}_i)D(n^{(1)},n^{(2)})},$$

$$(\mathbf{n}^{(1)},\mathbf{n}^{(2)}) \mapsto (\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)}+\mathbf{e}_i) \text{ w. prob. } \frac{cn_i^{(1)}p_2(\mathbf{e}_i|\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)})}{p_1(\mathbf{e}_i|\mathbf{n}^{(1)}-\mathbf{e}_i,\mathbf{n}^{(2)})D(n^{(1)},n^{(2)})},$$

$$(\mathbf{n}^{(1)},\mathbf{n}^{(2)}) \mapsto (\mathbf{n}^{(1)}+\mathbf{e}_i,\mathbf{n}^{(2)}-\mathbf{e}_i) \text{ w. prob. } \frac{Kcn_i^{(2)}p_1(\mathbf{e}_i|\mathbf{n}^{(1)},\mathbf{n}^{(2)}-\mathbf{e}_i)}{p_2(\mathbf{e}_i|\mathbf{n}^{(1)},\mathbf{n}^{(2)}-\mathbf{e}_i)D(n^{(1)},n^{(2)})},$$

for $i,j \in \{1,2\}$.

It remains to specify an approximation for the conditional sampling distributions $p_i(\cdot|\cdot)$. This was done for `K` and `W` in [SD00], and for `TI` in [DG04]. A natural approach would be to modify the generator-based method of [DG04] for `S`, but the resulting conditional sampling distribution vanishes for types which are present in the seed bank, but not in the active population, because mergers are blocked in the seed bank. The trunk ancestry method of [PS10] fails for the same reason. Instead, we suggest pooling the two populations and averaging the rates of mergers and mutations. More precisely, let $\hat{p}_{SD}(\mathbf{e}_i|\mathbf{n};u)$ be the approximate conditional sampling distribution of [SD00] for `K` with mutation rate $u$, and define

$$\hat{p}(\mathbf{e}_i|\mathbf{n}^{(1)},\mathbf{n}^{(2)}) := \hat{p}_{SD}(\mathbf{e}_i|\mathbf{n}^{(1)}+\mathbf{n}^{(2)};u+u'/K),$$

where the mutation rate has been obtained as the ratio of the average mutation rate, $uK/(K+1)+u'/(K+1)$ and the average merger rate $K/(K+1)$.

# References


[AR09] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.*, 37:697–725, 2009.

[ASS07] U. Arunyawat, W. Stephan, and T. Städler. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol. Biol. and Evol.*, 24(10):2310–2322, 2007.

[BB08] M. Birkner and J. Blath. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.*, 57(3):435–465, 2008.

[BBGW19] J. Blath, E. Buzzoni, A. González Casanova, and M. Wilke Berenguer. Structural properties of the seed bank and the two island diffusion. *J. Math. Biol.*, 79(1):369–392, 2019.

[BGE+15] J. Blath, A. González Casanova, B. Eldon, N. Kurt, and M. Wilke Berenguer. Genetic variability under the seedbank coalescent. *Genetics*, 200(3):921–934, 2015.

[BGKS13] J. Blath, A. González Casanova, N. Kurt, and D. Spanò. The ancestral process of long-range seed bank models. *J. Appl. Probab.*, 50(3):741–759, 2013.

[BGKW16] J. Blath, A. González Casanova, N. Kurt, and M. Wilke Berenguer. A new coalescent for seed-bank models. *Ann. Appl. Probab.*, 26(2):857–891, 2016.

[BGKW19] J. Blath, A. González Casanova, N. Kurt, and M. Wilke Berenguer. The seed bank coalescent with simultaneous switching. *arXiv:1812.03783*, 2019.

[BGR09] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J. R. Stat. Soc. B*, 65:3–55, 2009.

[DG04] M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories II: Subdivided population models. *Adv. in Appl. Probab.*, 36(2):434–454, 2004.

[dHP17] F. den Hollander and G. Pederzani. Multi-colony Wright-Fisher with seed-bank. *Indag. Math. (N.S.)*, 28(3):637–669, 2017.

[EBBF15] B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

[EK86] S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

[Eth11] A. Etheridge. *Some mathematical models from population genetics*, volume 2012 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011.

[HC97] D. L. Hartl and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, MA, 1997.

[Her94] H. M. Herbots. *Stochastic models in population genetics: genealogical and genetic differentiation in structured populations*. PhD thesis, University of London, 1994.

[HL66] J. L. Hubby and R. C. Lewontin. A molecular approach to the study of genic heterozygosity in natural populations. I. the number of alleles at different loci in Drosophila Pseudoobscura. *Genetics*, 54(2):577–594, 1966.

[HMTŽ18] L. Heinrich, J. Müller, A. Tellier, and D. Živković. Effects of population- and seed bank size fluctuations on neutral evolution and efficacy of natural selection. *Theor. Popul. Biol.*, 123:45–69, 2018.

[HSJB19] A. Hobolth, A. Siri-Jégousse, and M. Bladt. Phase-type distributions in population genetics. *Theor. Popul. Biol.*, 127:16–32, 2019.

[Jen12] P. A. Jenkins. Stopping-time resampling and population genetic inference under the coalescent model. *Stat. Appl. Genet. Mol.*, 11:Article 9, 2012.

[Kin82] J. F. C. Kingman. The coalescent. *Stoch. Proc. Appl.*, 13:235–248, 1982.

[KKL01] I. Kaj, S. M. Krone, and M. Lascoux. Coalescent theory for seed bank models. *J. Appl. Probab.*, 38:285–300, 2001.

[KMTŽ17] B. Koopmann, J. Müller, A. Tellier, and D. Živković. Fisher-Wright model with deterministic seed bank and selection. *Theor. Popul. Biol.*, 114:29–39, 2017.

[LJ11] J. T. Lennon and S. E. Jones. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat. Rev. Microbiol.*, 9(2):119–130, 2011.

[Mau07] H. Maughan. Rates of molecular evolution in bacteria are relatively constant despite spore dormancy. *Evolution*, 61:280–288, 2007.

[MRG+16] O. Mazet, W. Rodríguez, S. Grusea, S. Boitard, and L. Chikhi. On the importance of being structured: instantaneous coalescence rates and human evolution — lessons for ancestral population size inference? *Heredity*, 116:362–371, 2016.

[Not90] M. Notohara. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.*, 29(1):59–75, 1990.

[PS10] J. S. Paul and Y. S. Song. A principled approach to deriving approximate conditional sampling distributions in population genetic models with recombination. *Genetics*, 186:321–338, 2010.

[RMG+18] W. Rodríguez, O. Mazet, S. Grusea, A. Arredondo, J. M. Corujo, S. Boitard, and L. Chikhi. The IIRC and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, 121:663–678, 2018.

[Rou04] F. Rousset. *Genetic structure and selection in subdivided populations*. Princeton University Press, 2004.

[SAMT19] T. Sellinger, D. Abu Awad, M. Möst, and A. Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *bioRxiv 701185*, 2019.

[SBB13] M. Steinruecken, M. Birkner, and J. Blath. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theor Pop Biol*, 87:15–24, 2013.

[SD00] M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. R. Statist. Soc. B*, 62(4):605–655, 2000.

[SL18] W. R. Shoemaker and J. T. Lennon. Evolution with a seed bank: the population genetic consequences of microbial dormancy. *Evol Appl.*, 11(1):60–75, 2018.

[STRR15] C. Sherlock, A. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Stat.*, 43:238–275, 2015.

[TLL+11] A. Tellier, S. J. Y. Laurent, H. Lainer, P. Pavlidis, and W. Stephan. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proc. Natl. Acad. Sci. U.S.A.*, 108(41):17052–17057, 2011.

[VRP00] R. H. Vreeland, W. D. Rosenzweig, and D. W. Powers. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, 407:897–900, 2000.

[Wak09] J. Wakeley. *Coalescent Theory: An Introduction*. Coalescent theory: an introduction. Roberts & Company Publishers, Greenwood Village, 2009.

[Wri51] S. Wright. The genetical structure of populations. *Ann. Eugen.*, 15(1):323–354, 1951.

[ŽT12] D. Živković and A. Tellier. Germ banks affect the inference of past demographic events. *Mol. Ecol.*, 21(22):5434–5446, 2012.